impossible worlds

FRANCESCO BERTO | MARK JAGO

Impossible Worlds

Impossible Worlds

Francesco Berto and Mark Jago

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford, OX2 6DP,
United Kingdom

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide. Oxford is a registered trade mark of
Oxford University Press in the UK and in certain other countries

First Edition published in 2019

Impression: 1

Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America

British Library Cataloguing in Publication Data

Data available

Library of Congress Control Number: 2019936333

ISBN 978–0–19–881279–1

Printed and bound by
CPI Group (UK) Ltd, Croydon, CR0 4YY

For Anna and Valeria

Contents

Introduction

The latter half of the twentieth century witnessed an 'intensional revolution': a great collective effort to analyse notions which are absolutely fundamental to our understanding of the world and of ourselves – from meaning and information to knowledge, belief, causation, essence, supervenience, conditionality, as well as nomological, metaphysical, and logical necessity – in terms of a single concept. This was the concept of a *possible world*: a way things could have been.

Possible worlds found applications in logic, metaphysics, semantics, game theory, information theory, artificial intelligence, the philosophy of mind and cognition. In 1986, in *On the Plurality of Worlds*, David Lewis called possible worlds 'a philosophers' paradise'. Whatever view one had on the kinds of things possible worlds are, there was widespread agreement on their being an indispensable theoretical tool.

That paradise has turned out to be full of problems. These have emerged in piecemeal fashion, as difficulties for this or that application of the possible worlds paradigm. It seems to us, however, that the difficulties revolve around a single issue. Most of those fundamental notions are *hyperintensional*: they require distinctions the standard possible worlds apparatus cannot easily make.

When we set out to write about impossible worlds – ways things could *not* have been – we decided to set our narrative against the background of an envisaged twenty-first century 'hyperintensional revolution'. A number of accounts have been developed, which

qualify as hyperintensional in some sense. They range from two-dimensional semantics (Chalmers 2006), to theories of aboutness (Yablo 2014), truthmaker semantics (Fine 2017), metaphysical grounding (Correia and Schnieder 2012), structured propositions (King 2011), transparent intensional logic (Duzi et al. 2010), and various non-classical logical approaches (Dunn and Restall 2002). How such theories, or families thereof, are connected to each other and how their relative merits can be assessed, are at present largely open questions. But whatever position impossible worlds take in this landscape, we believe that they will play a role in the revolution, and we felt the time was ripe for a book providing guidance through the burgeoning literature on the subject.

This book includes an opinionated introduction to theories and uses of impossible worlds. (A shorter and simplified presentation can be found in our 'Impossible Worlds' entry in the *Stanford Encyclopedia of Philosophy*.) We have our own preferences on the metaphysics of impossible worlds and the logical and philosophical applications they afford. We don't hide those preferences; but we have tried to provide fair accounts of the alternative views and to assess them in a balanced way.

The book also includes our own original proposals on a number of topics involving impossible worlds. Some of these have appeared previously in print, although often not in the form they appear here. We have drawn on material from Berto's papers 'Impossible Worlds and Propositions' (*The Philosophical Quarterly*, 2010), 'On Conceiving the Inconsistent' (*Proceedings of the Aristotelian Society*, 2014), 'Impossible Worlds and the Logic of Imagination' (*Erkenntnis*, 2017), 'Conceivability and Possibility: Some Dilemmas for Humeans' (with Tom Schoonen, *Synthese*, 2018), 'Truth in Fiction, Impossible Worlds, and Belief Revision' (with Chris Badura, *Australasian Journal of Philosophy*, 2018), 'Williamson on Counterpossibles' (with Rohan French, Graham Priest, and Dave Ripley, *Journal of Philosophical Logic*, 2018), and on Berto's book *Ontology and Metaontology* (with Matteo Plebani, Bloomsbury, 2015). We have drawn on material from Jago's papers 'Against Yagisawa's Modal Realism' (*Analysis*, 2013),

'The Content of Deduction' (*Journal of Philosophical Logic*, 2013), 'Recent Work in Relevant Logic' (*Analysis*, 2013), 'The Problem of Rational Knowledge' (*Erkenntnis*, 2013), and on Jago's book *The Impossible* (Oxford University Press, 2014). We are very grateful to all the editors and publishers for permission to use these works.

Outline of the Book

The book is divided into three parts. Part I deals with foundational issues. In Chapter 1, we survey a number of applications of possible worlds; find them all wanting; trace the problem back to hyperintensionality; and suggest that impossible worlds may help. We present various definitions of the notion of an impossible world from the literature. Such worlds make sense only if we can genuinely think about the impossibilities they represent. We argue that we can.

A central philosophical issue with worlds, possible or impossible, is how they represent what they represent. This is obviously connected to the problem of what kind of things they are. In Chapter 2, we discuss a number of different proposals. Perhaps impossible worlds are metaphysically different from possible worlds, and represent in a different way. Or perhaps they are metaphysically on a par with possible worlds. Impossible worlds may be taken as 'genuine' entities which, like Lewisian possible worlds, represent something as being an *F* by having a real *F* as a part. Or, they may be taken as non-existent objects. Or as abstract entities which, like the objects of general object theory, represent by encoding. Or they may be taken as primitive entities, with no questions asked on how they represent. Or maybe there are no such worlds: we should take a fictionalist stance, and just make believe that there are.

We argue that all such views face difficulties, and conclude that some *ersatz* approach fares the best. After characterizing the notion of an ersatz world in general terms, we notice that there are different ways to specify the view. We delve into the options in Chapter 3. Ersatz possible worlds can be understood as maximal

states of affairs, maximal properties, recombinations of bits of actuality, maps, or things built out of propositions or sentences. We argue that, when extended to impossible worlds, most of these approaches face issues: they either collapse into other views, or are not general enough to accommodate all the impossibilites we may want. We conclude that linguistic ersatzism, which views worlds as constructions from sentences of a 'worldmaking' language, is the most promising metaphysics of impossible worlds. We close Chapter 3 by discussing a problem it, together with the other variants of ersatzism, faces: the problem of aliens.

Parts II and III of the book are about the logical and philosophical applications of impossible worlds. The boundary between logic and philosophy is to some extent arbitrary, as is our partition of the topics. Part II covers epistemic, doxastic, and various non-classical logics. Part III covers applications connected to issues in mainstream epistemology, information theory, the philosophy of fiction, and topics in semantics and the philosophy of language. But Part II is not completely free from philosophical discussion and Part III is not completely devoid of formalism, although we have tried to keep technicalities under control throughout the book.

In Part II, Chapter 4, we introduce normal modal logics and their frame semantics. We then show how impossible worlds can be used to model *non-normal* modal logics, in which the Rule of Necessitation is not valid. We discuss further uses, involving *non-adjunctiveness* and *non-primeness*. Two general patterns emerge in these applications. Firstly, impossible worlds are generally understood as 'logic violators': worlds where some logical law fails. Secondly, in semantics of this kind truth conditions are often not spelled out uniformly: they differ between possible and impossible worlds. This raises a philosophical problem, whose discussion is postponed until Part III: what of compositionality, a basic requirement for a theory of meaning?

Chapter 5 deals with applications in epistemic and doxastic logic. Here the central topic is the problem of logical omniscience. The standard view models agents as knowing or believing all logical

truths and all logical consequences of what they know or believe. We discuss some approaches to avoiding this consequence which don't use impossible worlds, and find them wanting. A naïve impossible worlds approach can easily deliver a view which avoids this problem. But it faces a deeper problem of *bounded rationality*: how should the accessible impossible worlds be constrained, so as to model a moderately rational though not logically omniscient agent? We argue that closing worlds under a weaker-than-classical logic won't help. We also critically discuss a dynamic approach using impossible worlds, on which epistemic states evolve gradually towards closure.

Chapter 6 deals with the role impossible worlds play in the semantics of relevant logics. These are non-classical logics that aim to avoid the paradoxes of the material and strict conditional. The mainstream semantics here includes non-normal points of evaluation, which are naturally interpreted as impossible worlds. The discussion has revolved around making sense of the truth conditions for the relevant conditional and negation. We discuss information-theoretic interpretations of impossible worlds in this setting, and raise some issues. We also discuss interpretations guided by general views on conditionality and an interpretation in terms of truthmaking.

Chapter 7 presents an application of impossible worlds to modelling acts of imagination. We focus on a semantics for hyperintensional operators capturing a kind of mental simulation. We discuss a number of plausible constraints on such operators, including non-monotonicity, non-primeness, and a 'Principle of Imaginative Equivalents' that limits the hyperintensional anarchy of imagining.

In Part III, Chapter 8 revolves around a very general philosophical issue: is hyperintensionality a genuine phenomenon? Or is it a feature to be explained away, and which therefore does not require us to amend the standard possible worlds apparatus? We consider arguments for the latter view, and find them unsuccessful. We then focus on a general notion of hyperintensional content, and discuss two issues concerning it. Firstly, any hyperintensional theory of content must address the problem of granularity: how fine-grained must the relevant hyperintensional distinctions be? Secondly, we return to the

issue, flagged in Chapter 4, of non-uniform truth conditions, which raises a compositionality objection for theories of content. We argue that impossible worlds accounts can deliver a fully compositional theory of content.

Chapter 9 is about information, which we conceptualize semantically, in terms of ruling out scenarios. We argue that Frege's puzzle of informative identities, and the informativeness of logical inferences, can be accounted for hyperintensionally, using impossible worlds. In our favourite analysis, it may be indeterminate whether a given logical inference is informative. We also sketch an analysis of informative content in terms of what is said by a speaker making an utterance.

Chapter 10 deals with epistemic and doxastic contents. Here we focus on how to model a realistic cognitive agent, striking a balance between the implausible extremes of logical omniscience and complete logical ignorance. This is the problem of bounded rationality, flagged in Chapter 5. The belief states of such an agent seem to be closed under 'easy', trivial logical consequence, but not under full logical consequence. Yet the former seems to imply the latter. Our solution is that, while some trivial closure principle must fail in a belief state, it is indeterminate just where any such failure occurs. We give formal models of belief states so structured. These entail that nobody genuinely believes an outright contradiction. We close the chapter discussing the issue of people who claim they do.

Chapter 11, written with Chris Badura, applies impossible worlds to the analysis of truth in fiction and the metaphysics of fictional objects. We show that inconsistent fictions are naturally handled via a space of worlds including impossible worlds, and that truth in fiction can be understood as a kind of simulated belief revision over such a space, triggered by the fiction's explicit content. We then discuss fictionalist, realist, and Meinongian accounts of fictional characters, their problems, and their relative merits. We show how impossible worlds can help to improve on some of these accounts.

Chapter 12, written with Rohan French, Graham Priest, and Dave Ripley, is about counterfactuality. The starting point here is the intuitive view that counterpossibles – counterfactual conditionals

with impossible antecedents – are not all vacuously true, independently of the truth value of the consequent. We discuss objections to the effect that this intuition should be explained away, and find them unconvincing. We then offer a non-vacuist semantics for counterpossibles that resorts to impossible worlds. This triggers a discussion of the so-called 'Strangeness of Impossibility Condition' (SIC). This relates to the idea that some pairs of worlds are closer to one another than others, and that we evaluate counterfactuals by considering the closest worlds. The (SIC), then, holds that, for any given possible world, any impossible world is further away from it than any possible world is. In the semantics, the substitutivity of rigidly coreferential terms fails in counterfactual contexts. This is arguably a problem. Another objection revolves around making sense of arguments by *reductio ad absurdum* in mathematical practice. We argue that both objections can be met.

Acknowledgements

Versions of papers relevant to the book have been presented by us in a number of workshops, seminars, and conferences in Australia, the Czech Republic, France, Germany, Italy, Japan, the Netherlands, Slovakia, Sweden, Switzerland, the UK, and the US. Those who asked good questions and came up with useful comments include Jc Beall, Thomas Brouwer, Colin Caret, Roberto Ciuni, Aaron Cotnoir, Franca D'Agostini, John Divers, Mike Dunn, Richard Dietz, Catarina Dutilh Novaes, Rohan French, Emmanuel Genot, Vittorio Hoesle, Justine Jacot, Jonathan Jenkins-Ichikawa, Ira Kiourti, Barteld Kooi, Ernie Lepore, Tito Magri, Diego Marconi, Friederike Moltmann, Bence Nanay, Daniel Nolan, Lucy O'Brian, Hitoshi Omori, Francesco Orilia, Michele Paolini Paoletti, Matteo Plebani, Shahid Rahman, Stephen Read, Manuel Rebuschi, Greg Restall, Dave Ripley, Maciej Sendlak, Sebastian Sequoiah-Grayson, Jeroen Smid, Matthew Soteriou, Bob Stalnaker, Jacopo Tagliabue, Stephan Torre, Giuliano Torrengo, Martin Vacek, Mark Van Atten, Alberto Voltolini, Heinrich Wansing,

Zach Weber, David Wiggins, Crispin Wright, and doubtless many others.

Special thanks go to Chris Badura, JC Bjerring, Jorge Ferreira, Ed Mares, Graham Priest, Tom Schoonen, and the referees of Oxford University Press, who read (various parts of) the manuscript and provided helpful comments. Any remaining errors after their scanning are entirely their fault.

Part I

Impossibilities

1

From Possible to Impossible Worlds

1.1 Worlds as Ways

Things might have been otherwise. David Bowie may still have been with us, the sun may have been shining on Nottingham, and the Axis powers may have won the Second World War. Such alternative ways we call *possible worlds*. Each possible world is a way things could have been. (This initial characterization says nothing of what possible worlds are, metaphysically speaking. That's the topic of Chapters 2 and 3.) The actual world is the most general and comprehensive way in which things in fact are. In the actual world, the Nazis lost the Second World War, the sky one of us sees from his office in Nottingham is cloudy, and David Bowie died at the beginning of 2016.

Ways things could have been can resemble the way things actually are. A world where the Axis powers won the Second World War is still a world where there was a war in which the Nazis fought, though with a different outcome from the actual world. Some possible worlds involve only small changes from ours: think of a world exactly like the actual one, except that you are one inch taller. Others are very different: think of one where the laws of biology and physics are turned upside down, so that you can be born twice, or travel faster than the speed of light. As we will see, the idea that it makes sense to speak of relations of similarity between possible worlds is important for some applications.

Possible worlds have a *vast* array of applications. According to some, this is the main reason for accepting them: 'it may be that the best philosophical defence that one can give for possible worlds is to use them in the development of substantive theory' (Stalnaker 1991, 141). Since the late twentieth century rejection of the Quinean and Davidsonian idea that only extensional concepts should be allowed in serious philosophical inquiry, the notion of possible world has become ubiquitous in contemporary philosophy. It plays a key role in most branches of the discipline, ranging from logic to metaphysics and ontology, the philosophy of mind, the philosophy of information, moral and political philosophy, and aesthetics. But it has been used also outside of philosophy, in fields that range from the semantics of natural language to game theory, artificial intelligence, and cognitive science. We start with an overview of these applications. (Parts of the following section draw on Berto and Plebani 2015, chapter 11.)

1.2 Possible Worlds at Work

Possibility and Necessity

Perhaps the most typical application of possible worlds is in modal logic. This is, first of all, the logic of expressions like 'necessarily', 'possibly', 'contingently'. Such expressions are used in two different ways. A first use consists in qualifying the truth of a sentence, or of the proposition expressed by the sentence:

(1.1) It is necessary that $7 + 5 = 12$.

(1.2) It is possible that Scotland leaves the UK.

(1.3) Possibly, Anna wins the music contest.

(1.4) Necessarily, Valeria is human.

Modalities of this kind are called *de dicto*. Expressions like 'necessarily' or 'it is possible that', or the concepts they express, are attached

to *dicta*, that is, to pieces of language, or language-like entities, such as sentences or propositions. They express the way that sentence, or proposition, bears its truth value. Thus, according to (1.1), that seven plus five is twelve is necessarily true, and according to (1.3), that Anna wins the music contest is possibly true.

Modal expressions can also be used to qualify the features of objects:

(1.1a) Seven is necessarily an odd number.

(1.2a) Scotland is such that it could leave the UK.

(1.3a) Anna is a possible winner of the music contest.

(1.4a) Valeria is necessarily human.

Modalities of this kind are called *de re*, for the modals are used here to express the way in which a thing, a *res*, has some feature. Thus, according to (1.1a) and (1.4a), seven has the property of being odd, and Valeria that of being human, in a necessary way.

Contemporary logicians and philosophers follow Leibniz's insight that the necessary is what holds *no matter what*, in any way things could have been: that is, in all possible worlds. What is possible, on the other hand, is what holds at some possible world. What is contingent is what holds at some, but not all, possible worlds. Necessity and possibility are thus interpreted as quantifications over possible worlds. Using '$\Box$' for 'necessarily', '$\Diamond$' for 'possibly', 'iff' for 'if and only if', and letting W be the total set of possible worlds, we get:

'$\Box A$' is true at world w iff A is true at all worlds $w_1 \in W$

'$\Diamond A$' is true at world w iff A is true at some world $w_1 \in W$

(The notation '$w_1 \in W$' here means that w_1 is a member of the set W. It's a way of expressing that w_1 is a possible world.)

The two notions $\Box$ and $\Diamond$ are *duals* of one other, just as the universal and particular quantifiers, $\forall x$ and $\exists x$, are of one another. Each modal can be defined via the other and negation. That it is necessarily the

case that A means that it is not possible that $\neg A$ ('not-A'). And that it is possible that A means that it is not necessary that $\neg A$.

Necessity and possibility are highly ambiguous notions. (For a taxonomy, see chapter 1 of Divers 2002.) Although there is no universal consensus on this, many philosophers adopt three kinds of *absolute necessity*, holding in all possible worlds unrestrictedly:

LOGICAL NECESSITY fixed by the laws of logic broadly conceived (e.g., that if A, then either A or B);

MATHEMATICAL NECESSITY fixed by mathematical truths (e.g., that $7 + 5 = 12$); and

METAPHYSICAL NECESSITY fixed by the identity and nature of things (e.g., that water is H_2O; that Valeria is a human being).

We will not get into the issue of whether one of these is reducible to another (e.g., the mathematical to the logical, as claimed by *logicists* in the philosophy of mathematics, including Dedekind (1901), Frege (1879), Peano (1889), and Russell (1903)).

We also talk of things being necessary, or impossible, only in a relative sense, or from a certain viewpoint. We are stuck in a traffic jam in Paris at 2 pm; our flight is leaving from De Gaulle airport at 2:10 pm. We moan: 'There's no way that we can make it to the airport in time'. What we mean is that, given the timing, the means of transport available, and the laws of physics of our world, it is impossible for us to reach the airport in time. It is not unrestrictedly, absolutely impossible: if we had *Star Trek*'s transporter, we could make it. But a *Star Trek* world in which one can be instantaneously disassembled into atoms and reassembled exactly with the same atomic structure in a different place is a world quite different from ours. One may doubt that such a world is even physically possible, that is, compatible with our laws of physics.

Other modal notions, thus, are naturally understood as restricted forms of necessity or possibility. Something can count as R-necessary, for some relativized modal notion R, even if it fails to hold at some

possible world or other. Accordingly, the corresponding modals are understood as restricted quantifiers over possible worlds. Thus *nomological necessity*, compliance with the laws of nature of the actual world or of the world under consideration, is often (no universal consensus here either) taken to be a relative or restricted necessity. It is biologically impossible but not absolutely impossible for a human being to jump one mile up in the air; it is physically impossible (if Einstein was right) but not absolutely impossible for a body to travel faster than the speed of light.

Propositions

Possible worlds are extremely important for theories of *representation*, both in language and thought, and have been used to analyse key notions from the philosophy of language. Many of these approaches build on Wittgenstein's insight that understanding the meaning of a sentence is grasping its truth conditions: 'to understand a proposition means to know what is the case if it is true' (Wittgenstein 1921/1922, §4.024). Montague (1970) and Stalnaker (1976a) have claimed that *propositions*, the meanings or contents expressed by sentences and the primary bearers of truth values, should be understood as sets of possible worlds. The proposition expressed in English by 'raccoons like to somersault' is, on that view, the set of possible worlds where raccoons like to somersault, precisely the same set of possible worlds making for the proposition expressed in Italian by 'ai procioni piace fare le capriole'.

Knowledge and Belief

Another notion analysed via possible worlds is *knowledge*. Following Hintikka (1962), knowledge has been characterized in terms of what is true throughout all the ways things could be, for all the agent in question knows. On this approach, the possible worlds accessible to an agent represent her *epistemic possibilities*. Knowledge can then be treated as another restricted quantifier over possible worlds. If K stands for a given agent's state of knowledge, and R is a binary

accessibility relation on the space of worlds W, the Hintikka-style characterization goes thus:

(H) KA is true at w iff A is true at all w_1 such that Rww_1

This thought is at the core of contemporary epistemic logic (see, e.g., Blackburn et al. 2002, Fagin et al. 1995, Van Benthem 2003). But several research programs in mainstream epistemology also rely on a similar viewpoint. Dretske's *relevant alternatives* approach takes knowledge as 'an evidential state in which all relevant alternatives (to what is known) are eliminated' (Dretske 1981, 367). Lewis (1996) discusses a similar approach. Alternatives here work similarly to possible worlds, and the uneliminated relevant alternatives work similarly to accessible worlds.

Necessity (whether logical, mathematical, metaphysical, or nomological) and knowledge share the feature of being *factive*: what is necessary, and what is known, is true. Factivity can be expressed by claiming that the actual world must always be one of the (accessible) possible ones, with respect to the relevant kind of possibility. Unlike other factive modalities, though, knowledge is an *intentional* state: a state of the mind directed towards a certain content. There are also non-factive intentional states, including belief, desire, fear, hope, and imagination. These have also been understood using restricted quantifiers over possible worlds, where the accessible worlds are the ones where things are as the agent believes (imagines, etc.) them to be. (Fagin et al. 1995 is a comprehensive guide through epistemic and doxastic logics. Niiniluoto (1985) and Wansing (2017) each discuss the application to imagination; Berto (2018) gives a semantics for imagination using an enriched possible worlds approach.)

Knowing (or believing, imagining, etc.) that A is often taken to be a mental state whose content is the proposition expressed by A. As well as being the primary bearers of truth values, thus, propositions have been understood as the content of (*de dicto*) intentional states: they are what is known, believed, feared, or imagined when one knows, believes, fears, or imagines something (*de dicto*). Just as different

sentences like 'raccoons like to somersault' and 'ai procioni piace fare le capriole' can share the same content, so can different people's mental states share the same content: John believes, and Mary fears, that Marine will win the elections. There is some doubt that one and the same kind of entity can cover both the role of primary truth bearers and the role of targets of *de dicto* intentional states (Jago 2018b, Lewis 1986b). Nevertheless, possible worlds stake a good claim at giving a unified account of a broad range of linguistic and mental contents.

Information

Information is connected to knowledge, or potential knowledge. If a sentence or proposition is informative, then one can come to know that information (say, by hearing the sentence uttered truthfully by a trusted speaker). Something may be informative even if no one yet knows it, however. We might think of information as embodying potential knowledge for some suitable cognitive agent in the right circumstances. If we analyse knowledge in terms of possible worlds, we should expect a similar approach to information to be available.

According to the *Bar-Hillel-Carnap theory of information* (Bar-Hillel and Carnap 1953, Bar-Hillel 1964), the informative job of a sentence *A* consists in partitioning the totality of possible worlds into those where *A* is true and those where it is false. We may identify the information with the partitioning function, which in effect says 'yes' to some possible worlds and 'no' to all the others. Or we might identify the information with the set of 'yes' worlds. (Mathematically speaking, the former is the *characteristic function* of the latter set. The two approaches are, in a straightforward mathematical sense, equivalent.)

You might notice the similarity to the possible worlds account of propositions: both are treated as sets of possible worlds. That's no coincidence. On this approach, the information contained in a sentence (in a context) is precisely the proposition expressed by an utterance of it (in that context).

This gives us a static notion of information, as something that's possessed by a sentence or proposition. The possible worlds approach also allows us to account for a dynamic notion of information, of *becoming informed* of such-and-such. When a cognitive agent gains the information (and, let's suppose, thereby learns) that raccoons like to somersault, we can model this in terms of ruling out the worlds where it is not the case that raccoons like to somersault. (Perhaps the raccoons of those worlds have different tastes; perhaps there are no raccoons there at all.) By 'rule out', we don't mean that the agent thereby treats those worlds as impossibilities. Rather, she rules them out as contenders for actuality: the ways things are, for all she knows.

Indicative Conditionals

Conditionality may also be dealt with using possible worlds. Many philosophers and logicians are unsatisfied with the material conditional, '$\supset$', taken as the operator given by the usual (two-valued) truth table: '$A \supset B$' is false when A is true and B false, true otherwise. This delivers two inferences which have sometimes be called 'paradoxes of the material conditional' (Anderson and Belnap 1975, MacColl 1908, Routley et al. 1982):

(1.5) If $\neg A$, then $A \supset B$

(1.6) If B, then $A \supset B$

If we try to understand the English indicative conditional 'if … then' in terms of $\supset$, many seemingly false conditionals will come out true, just because their antecedent is false, or their consequent true:

(1.7) If Obama is Canadian, then the Moon is made of green cheese.

(1.8) If Strasbourg is in Germany, then Obama is American.

One reason to reject these is that there is no relevant connection between their antecedent and consequent: what's Obama's nationality got to do with the constitution of the moon, or the location of

European cities? Another reason to reject (1.7) and (1.8) is that any connection between their antecedent and consequent seems far too contingent. Even if Obama were Canadian, the moon would not be a giant cheeseball.

This suggests an alternative conception of the conditional on which, for 'if A, then B' to be true, it *cannot* be the case that A is true while B is false. This analysis gives us the *strict conditional*, '$\neg\!\!\exists$'. '$A \neg\!\!\exists B$' is true just in case there is no possible world in which A is true but B is not. The strict conditional is the necessitation of the material conditional: $A \neg\!\!\exists B$ is understood as $\Box(A \supset B)$. It's easy to see that (1.5) and (1.6) are invalid when we replace '$\supset$' with '$\neg\!\!\exists$'. (Whether this move really avoids the worries is something we'll come back to below, in §1.3 and Chapter 6).

Counterfactual Conditionals

Possible worlds have also been used to give a semantics for *counterfactual* conditionals. These are conditionals of the form 'if it were (or, had been) the case that A, then it would be (or, have been) the case that B', symbolized as '$A \mathrel{\Box\!\!\rightarrow} B$'. Counterfactuals are so-called because, in a typical use, they have a false antecedent, *contra factum*. In explaining why kangaroos have tails, for example, we might say 'well, if kangaroos had no tails, they would topple over' (Lewis 1973b).

(Many philosophers use 'counterfactual' for any conditional of the form, 'if it were …, then it would be …', even if the antecedent is true (Bennett 2003, Lewis 1973b, Williamson 2007). Others prefer to talk of 'subjunctive conditionals'. We'll will stick to the standard 'counterfactuals' terminology for all such conditionals.)

Counterfactuals are extremely important in our cognitive lives. We conceive counterfactual alternatives to reality in order to explore what would or would not happen, were those alternatives realized. Would John not have been injured, had he avoided crossing the road? They are also important in understanding history (Nolan 2016b): what if Hitler had had the A-bomb in 1944? They may help us to understand the concept of causation better (Lewis 1973a, Paul 2004;

see Paul 2009, Paul and Hall 2013 for in-depth discussion). So it's important to give a semantic analysis of counterfactuals.

How? That counterfactuals must be modal conditionals can be argued by comparing them to the corresponding indicative conditionals:

(1.9) If Kate Bush didn't write 'The Kick Inside', someone else did.

(1.10) If Kate Bush hadn't written 'The Kick Inside', someone else would have.

These have the same antecedent and consequent (in different moods), but different truth values. (1.9) seems true. We know that someone wrote 'The Kick Inside', so if it wasn't Kate Bush, it must have been someone else. By contrast, (1.10) seems false: 'The Kick Inside' might never have been written, if it hadn't been for Kate Bush. So even if one insists that (1.9) be taken as a material conditional, (1.10) seems to be of a different kind. The difference in mood between (1.9) and (1.10) has been understood as getting us to evaluate (1.10) by looking at alternative ways things could have been, that is, at alternative possible worlds.

Which worlds? The mainstream treatment of counterfactuals, due to Stalnaker (1968) and Lewis (1973b), says that we should evaluate 'if kangaroos had no tails, they would topple over' by looking to the *closest* possible worlds where kangaroos have no tails. We then see whether kangaroos topple over there. *Closeness* between worlds is understood as involving (contextually determined) similarity in the relevant respects. So evaluating a counterfactual will typically involve the *minimal change* (with respect to the world of evaluation) required to verify the antecedent. We disregard worlds where kangaroos have no tails but help themselves with crutches, or have evolved wings. Overall: '$A \mathrel{\Box\!\!\rightarrow} B$' is true (at world w) iff the closest(-to-w) possible A-worlds are B-worlds.

(What if several possible worlds tie for closeness? Do we require *all* closest A-worlds to be B-worlds? Or some? Or most? What if the A-worlds get forever closer and closer to ours, with none being the

closest? These are tricky questions: Kratzer (1981), Lewis (1973b, 1981), and Nute (1975) discuss them in detail. We won't get into them here.)

Possible worlds have also been used in the analyses of *essence* and *de re modality* (Lewis 1986b), and of *metaphysical dependence* and *supervenience* (Bennett 2004, Davidson 1970). Many physicalist philosophers of mind, including Horgan (1982, 1993), Kim (1982, 1993), and Lewis (1983), express their commitment to physicalism about mental states in terms of supervenience, cashed out in terms of possible worlds. But, for reasons we won't go into here, we don't think a worlds-based approach is the best way to capture notions of essence or dependence. (We're largely persuaded by Fine's (1994) arguments.) So we won't discuss these applications any further.

Possible worlds are a success story of philosophical theorizing. Still, most of the accounts using them, which we have just sketched, face issues. The umbrella under which many of these can be gathered is the concept of *hyperintensionality*, to which we now turn.

1.3 The Problem(s) of Hyperintensionality

Hyperintensionality can be characterized as a feature of concepts. A concept is hyperintensional when it draws a distinction between necessarily equivalent contents, where the relevant necessity is unrestricted: logical, mathematical, or metaphysical, if we stick to the threefold distinction mentioned above. If the relevant concept is expressed by an operator $\mathscr{H}$, then $\mathscr{H}$ is hyperintensional when $\mathscr{H}A$ and $\mathscr{H}B$ can differ in truth value, in the face of A and B's being necessarily (logically, mathematically, or metaphysically) equivalent.

(Cresswell (1975) originally defined 'hyperintensional' to pick out a position in a sentence in which logical equivalents cannot be replaced *salva veritate*. But, as Nolan (2014, 151) notes, it is now common to use the term more broadly, with 'necessary equivalence' in place of 'logical equivalence'.)

This characterization of hyperintensionality is a contrastive one. It tells us that a concept or operator is hyperintensional when it is more fine-grained than intensional or (normal) modal concepts or operators, marking a distinction invisible to the latter. It does not yet provide us with a full-fledged characterization of hyperintensionality, and it says nothing about 'just "how hyper" hyperintensions are' (Jespersen and Duzi 2015, 527). Different hyperintensional notions may display different degrees of fine-grainedness. We discuss this key issue in Chapter 8.

The problems we are about to examine affect the possible worlds accounts introduced in §1.2. The problems emerged over the last few decades in piecemeal fashion. But a single issue underlies them all: they are hyperintensional notions, making distinctions more fine-grained than the standard possible worlds approach can easily model.

Propositions: Triviality

If we take propositions, the meanings or content of sentences, as sets of possible worlds, then necessarily equivalent propositions are one and the same proposition: possible worlds never disagree on necessarily equivalent sentences. Assuming again that mathematical and logical necessity are unrestricted, 'if Obama is human, then Obama is human' and '7 + 5 = 12' are true in the same possible worlds: all of them. So they express the same proposition, viz., the total set of worlds.

This seems wrong: the sentences should have different meanings. They speak of different things: only one is about Obama. We have a dual problem with sentences that cannot be true, like 'Obama is both human and not human' and '7+5 = 13'. These would also express the same possible-worlds proposition: the empty set of possible worlds. This seems just as bad a result as the first: the two sentences have different meanings and are about very different things: the first is about Obama, the second is not.

This problem is particularly evident when we turn to the kinds of propositions typically expressed when we do metaphysics. Many metaphysical claims are such that, if they are true, or false, they are

necessarily so. This includes claims of modal metaphysics, such as statements about the very nature of possible worlds. If we say that possible worlds have such-and-such natures, then we seem committed to that claim being necessarily true. After all, if it were possibly false, then it would be false at some possible world, which seems to make little sense. (Just how powerful this argument is depends on how we take worlds to represent a particular state of affairs: we'll discuss this issue in chapters 2 and 3.)

Many other metaphysical claims seem to be necessary (if true at all): Hegel's doctrine of the Absolute *Geist*, Plato's view of ideas as purely intelligible forms, and Armstrong's claim that there are immanent multiply instantiated universals, do not seem to be contingent claims. Defenders and objectors of these views alike agree that these are distinct viewpoints, expressed by distinct propositions. If that's right, then each view corresponds either to the set of all worlds (if true), or to the empty set (if false). But since these are three distinct views, expressed through three distinct propositions, those propositions are not plain sets of possible worlds.

Here's a further puzzle to bring out the problem. Suppose Anna and Valeria are debating the nature of properties. Anna says (P) that they're transcendent Platonic universals, whereas Valeria says (I) that they're immanent universals. Each view corresponds either to the set of all worlds (if true), or to the empty set (if false). Suppose further than both Anna and Valeria believe that propositions are sets of possible world. Then Anna must accept that her claim is identical to the claim that $P \vee I$ (since she believes P is necessarily true, I is necessarily false, and hence that P and $P \vee I$ each correspond to the set of all worlds).

Similarly, Valeria must accept that her claim is identical to the claim that $P \vee I$ (since she believes that P is necessarily false, that I is necessarily true, and hence that I and $P \vee I$ each correspond to the set of all worlds). But if each accepts that their claim is identical to the claim that $P \vee I$, they must accept that their claims are identical, which neither will accept. If they are to have a serious debate about the nature of properties, therefore, they should reject

their beliefs that propositions are sets of possible worlds. Genuine, rational metaphysical debate is possible only on the assumption that propositions are not sets of possible worlds.

Knowledge and Belief: Logical Omniscience

Historically, one of the first manifestations of the hyperintensionality issue came from the modal treatment of epistemic and doxastic concepts. Here, the issue is *logical omniscience*: a cluster of closure conditions on knowledge and belief, which come as a spin-off of Hintikka's (1962) possible-worlds approach (§1.2). Perhaps the most important closure effects of Hintikka's clause (H) above are:

(C1) If KA and A entails B, then KB

(C2) If A is valid, then KA

(C3) It is not the case that: KA and $K\neg A$

and similarly for belief. (We will find a more comprehensive list of closure conditions in §5.1.)

(C1), often dubbed *Closure under entailment* or *Full omniscience*, says that one knows all the entailments or logical consequences of what one knows. The principle also applies to the possible-worlds semantics for belief: one believes all the logical consequences of what one believes. (C2), *Knowledge of all valid formulas*, says that one knows all the logical truth (and similarly for beliefs). When we define validity as entailment by the null set of premises, (C2) is a special case of (C1). (C3) guarantees *Consistency* of knowledge: one can never have inconsistent knowledge, and the corresponding principle says one can never have inconsistent beliefs.

These conditions follow directly from interpreting the relevant epistemic notions as restricted quantifiers over possible worlds. For instance, (C1) holds once we understand A's entailing B as the claim that B is true at all possible worlds (of all models of the epistemic logic at issue) where A is true. Then, if A is known (believed), it holds at all the epistemically accessible possible worlds. But if A entails B,

then *B* holds at all those worlds too, and so *B* is known (believed) as well. (C2) holds when we understand the logical validity of *A* as its holding in all possible worlds (of all models, etc.). Then, in particular, a valid *A* holds at all the epistemically accessible worlds, and so is automatically known (believed).

For applications in computer science, such principles are often taken as harmless (Fagin et al. 1995, chapter 9). However, it is generally admitted that they deliver implausibly idealized notions of knowledge and belief, having little to do with *human* cognition. Against (C1), for instance: we know basic arithmetic truths like Peano's postulates, and these entail (let us suppose) Goldbach's Conjecture; but we don't know whether Goldbach's Conjecture is true. Against (C2): Excluded Middle is (let us suppose) valid, but intuitionist logicians do not believe it, and so do not know it either.

As for (C3): there cannot be inconsistent knowledge, given that knowledge is factive and assuming there are no true contradictions. But real, finite, and fallible cognitive agents may well have inconsistent beliefs. They may even believe the relevant inconsistencies explicitly, and take themselves as justified in doing so (e.g., *dialetheists* believe that the Liar sentence is both true and false (Priest 1987)).

An answer one sometimes hears is that *K* in (H) expresses not knowledge or belief, but rather some derivative attitude, characterized in terms of knowledge or belief: what an agent is logically committed to, given what else she knows or believes. This leaves us in want of a logical account of *knowledge* and *belief* for real agents, as opposed to some conditional commitment. One may also question this account of epistemic or doxastic commitment. Is an intuitionistic logician really committed to Excluded Middle (given classical logic)? Are those of us with inconsistent beliefs – all of us! – really committed to everything being true, given that a contradiction classically implies every sentence?

Information: Triviality and Overload

The possible worlds Bar-Hillel-Carnap analysis of information has similar issues to the account of knowledge and belief. 'If Obama is

human, then Obama is human' and '$x^n + y^n = z^n$ has no integer solutions for $n > 2$' are both necessarily true. So there is no possible world ruled out by learning either. On the Bar-Hillel-Carnap analysis, neither are genuinely informative, and so neither are learnable. But while the former is easily deemed true by competent speakers of English, the truth of the latter is non-trivial in the extreme. For the latter is Fermat's Last Theorem, a proof of which took centuries to find. The first, by Andrew Wiles, was 130-something pages long.

The problem generalizes. A possible worlds analysis of information entails that *no* logical, mathematical, or metaphysical truth can be informative. It denies, in particular, the informativeness of any logical deduction or mathematical proof, and thus the epistemic value of devoting one's time to the study of mathematics or logic. But some deductions and proofs are obviously informative. This can depend on the fact that the conclusion has high syntactic or semantic complexity, but it need not be so. Fermat's Last Theorem is expressed by a sentence anyone with high school maths can understand. But recognition of its truth, via proof, is extremely complicated.

Even simple proofs, like short truth table calculations, can be informative. Students who have just mastered the truth table for the material conditional may be surprised to find out that Frege's Law, $(A \supset (B \supset C)) \supset ((A \supset B) \supset (A \supset C))$, is a tautology, or that for all A and B, either $A \supset B$ or $B \supset A$. It is part of the explanation of why they are surprised, that they acquire new information. It seems, then, that there is a legitimate notion of information whereby one *can* learn, or become informed of, a tautology.

Indicative Conditionals: Irrelevance

The possible worlds treatment of conditionality is not free from problems either. We have seen that the strict conditional '$A \dashv B$' is free from the paradoxes of the material conditional. But it has its own so-called 'paradoxes of the strict conditional':

(1.11) If $\neg\Diamond A$, then $A \dashv B$

(1.12) If $\Box B$, then $A \dashv\!\!\vDash B$

If B is true in all possible worlds, or A in none, then there is no possible world where A is true and B is false, so '$A \dashv\!\!\vDash B$' is true too. Interpreting the 'if …, then …' of English as the strict conditional, this makes many seemingly false conditionals true, just because their antecedent is impossible, or their consequent is necessary:

(1.13) If $5 + 7 = 13$, then Obama is Canadian.

(1.14) If Obama is American, then $5 + 7 = 12$.

These look bad because of the irrelevance phenomenon. There seems to be no connection between the antecedent and consequent: they are about wholly distinct things. Given the necessity of logical truth and the impossibility of logical falsity, we also get true strict conditionals whose consequent is a truth of logic, or whose antecedent is a falsity of logic, e.g., of the following form:

(1.15) $A \dashv\!\!\vDash (B \dashv\!\!\vDash B)$

(1.16) $A \dashv\!\!\vDash (B \vee \neg B)$

(1.17) $(A \wedge \neg A) \dashv\!\!\vDash B$

These also look bad due to irrelevance: what A is about may have nothing to do with what B is about. Take an instance of (1.16), 'if the Moon is made of green cheese, then either Nottingham is in Scotland, or not'. Does that sound correct? (We will come back to this kind of irrelevance phenomenon in Chapter 6.)

Counterfactual Conditionals: Counterpossibles

Counterfactuals with impossible antecedents are called *counterpossibles*. The Lewis-Stalnaker treatment of counterfactuals delivers *vacuism*: the view that all counterpossibles are vacuously true. If '$A \mathrel{\Box\!\!\rightarrow} B$' is true when all the closest A-worlds are B-worlds, and there are no A-worlds, then it comes out automatically true. (Just as

'all hobbits in this room are tiny' is true, trivially, given that there are no hobbits in the room.) So a counterfactual '$A \boxrightarrow B$' will be trivially true whenever its antecedent is impossible. To add insult to injury, the conditional with the same antecedent and negated consequent, '$A \boxrightarrow \neg B$', will also be trivially true.

Some philosophers believe that, appearances to the contrary notwithstanding, this is all right (we discuss this kind of view in Chapter 12). However, many – including Nolan (1997), Brogaard and Salerno (2013), Priest (2008), Krakauer (2012), Bjerring (2014), and Bernstein (2016) – think these results to be problematic. Nolan (1997) gives a nice example:

(1.18) If Hobbes had (secretly) squared the circle, sick children in the mountains of South America at the time would have cared.

(1.19) If Hobbes had (secretly) squared the circle, sick children in the mountains of South America at the time would not have cared.

Hobbes' squaring the circle would have made absolutely no difference to the suffering of those sick children. So (1.19) should come out true true for *this* reason, and not merely because there are no worlds verifying the antecedent. Similarly, (1.18) should come out false. Some counterpossibles are false; and where they are true, typically, they are not trivially so.

The problem connects in an obvious way to the triviality problem for possible worlds propositions. We often reason counterfactually in matters of mathematics, logic, and metaphysics. Vacuism about counterpossibles can hardly account for this. We make counterfactual suppositions in all of these areas, perhaps with the purpose of criticizing a theory by drawing unpalatable consequences. Such a practice is trivialized if counterpossibles are all true. Imagine Hilbert arguing against Brouwer that, if intuitionism were true, then much of standard mathematics would be lost, for we could not then resort to impredicative definitions. Vacuism makes any such claim devoid of dialectical content. For given vacuism (and classical logic), we

can also truthfully assert that, if intuitionism were true, nothing of standard mathematics is lost. Indeed, this claim would be equivalent to Hilbert's; yet it is what Hilbert wanted to deny.

Taken together, these problems provide a strong case against the possible worlds approach. It might even be that they provide reason to abandon all attempts to analyse these notions in terms of worlds (Fine 1975a, 2012a, 2019).

However, we needn't abandon a worlds-based analysis of these concepts. The problems we've just sketched show that we can't give good enough analyses using only *possible* worlds. We have to be more open minded. We can give good worlds-based analyses of these concepts, so long as the worlds in question include possible *and impossible* worlds. Using impossible worlds, we can solve a number of problems faced by possible worlds accounts of knowledge, belief, meaning, information, and conditionality.

1.4 Impossible Worlds

This book is concerned with worlds that are not possible, with 'possible' understood in an unrestricted sense. You might worry that impossible worlds are metaphysically weird, logically disreputable, or not really useful for this or that purpose; or that they lose some crucial benefit of possible worlds accounts. We'll argue that it ain't so. In Chapters 2 and 3, we'll show that impossible worlds can be metaphysically acceptable even if, as we'll discuss there, some attempts to make them metaphysically reputable fail. In Chapters 4–7, we'll show how impossible worlds have useful logical applications, and that they may, but need not, involve a departure from classical logic. We'll discuss philosophical applications of impossible worlds in Chapters 8–12. As we go, we'll address some of the objections that have been raised against impossible worlds, including, for example, that they don't allow for a compositional account of meaning (Chapter 8).

The possible worlds framework seems still to be a dominant conceptual framework of our time for philosophical theorizing. We'll argue that the impossible worlds framework constitutes a net theoretical gain. The late twentieth century saw an intensional revolution centred on the notion of possible world. The early twenty-first century is seeing what Nolan (2014) called a *hyperintensional revolution*. Impossible worlds are at home in this revolution.

They aren't the only theoretical tool that's been suggested for analysing hyperintensionality. An alternative is the *structured propositions* approach of King (1995, 1996, 2007), Soames (1985, 1987), and others. Another one comes from Pavel Tichy's *transparent intensional logic* (Duzi et al. 2010). Other recent approaches include Fine's *truthmaker semantics* (Fine 2012a, 2014, 2019) and Yablo (2014)'s work on *aboutness*, which enriches possible worlds semantics with divisions of the space of worlds itself. We shan't discuss these here in any detail. We've discussed structured propositions elsewhere Jago (2014a, 2015, 2017). Whatever its merits in accounts of content, it cannot claim to be a general account of hyperintensional logical or philosophical notions. Ripley (2012) compares the impossible worlds and structured propositions approaches to hyperintensionality, coming down forcefully on the side of the former.

(Truthmaker semantics is an exciting recent development. As a general philosophical approach to hyperintensionality, it is at present underdeveloped, but has great potential. We don't think it can be a general approach to hyperintensional notions. It seems that any truthmaker involving James Newell Osterberg is thereby a truthmaker involving Iggy Pop, since Osterberg is Iggy Pop. But one can believe that Iggy Pop co-wrote David Bowie's 'China Girl' without believing that Osterberg did. So we don't see how epistemic or doxastic contents could be modelled using truthmaker semantics.)

Our aim is to investigate, develop, and defend the way of impossible worlds. Of the other ways, we won't say much. (For an assessment of the relative merits of the structured propositions approach, aboutness à la Yablo, truthmaking à la Fine, and impossible worlds, in the treatment of hyperintensionality, see Gioulatou (2016).) So let us turn

to the obvious question: what's an impossible world? (The following material draws on Berto and Jago 2018.)

A look at the literature on impossible worlds (which is rapidly growing: see Nolan 2013 for a survey) presents us with a number of different definitions. These can be reduced to four main ways of treating impossible worlds, ordered from the more to the less general:

IMPOSSIBLE WAYS: Just as possible worlds are characterized as ways things could have been, so are impossible worlds often characterized as ways things could *not* have been. The initial insight is that not everything is possible. Some things just (absolutely) cannot happen. Anything that just can't happen must be an absolute impossibility; and these ways the world just couldn't be are impossible worlds. Beall and van Fraassen (2003), Restall (1997),Salmon (1984), and Yagisawa (1988) think of impossible worlds in this way.

LOGIC VIOLATORS: Another definition has it that impossible worlds are worlds where the laws of logic fail. This approach depends on what we take the laws of logic to be. Given some logic L, an impossible world with respect to the L-laws is one in which some of those laws fail to hold (Priest 2008, chapter 9). An impossible world in this second sense will also be impossible in the first sense, so long as the logic L in question is no stronger than the logic governing logical possibility. But for dialethists or intuitionists, a world violating containing contradictions, or failing excluded middle, won't count as a way the world couldn't be, and so won't count as an impossible world in the first sense. Whichever logic is operative, there are worlds which count as impossible in the first but not in this second sense. If the Continuum Hypothesis of set theory is true (and, logicists are wrong!), some world where the Continuum Hypothesis fails may well be impossible in the first sense without thereby violating any law of the chosen logic.

CLASSICAL LOGIC VIOLATORS: Another definition has it that impossible worlds are worlds where the laws of classical logic fail (Priest

1997a). This definition gives the same results as the previous one if we take the laws of logic to be the classical ones, but not otherwise. A world complying with intuitionistic logic, but where instances of Excluded Middle fail, will be impossible in this third sense.

CONTRADICTION-REALIZERS: A still narrower definition has it that an impossible world is one where sentences of the form A and $\neg A$ hold, against the Law of Non-Contradiction (Lycan 1994). Impossible worlds of the fourth kind will be impossible in the third sense, since they thereby violate classical logic. But not vice versa: an intuitionistic world will have the Law of Non-Contradiction hold unrestrictedly, and so will be impossible in the third, but not the fourth, sense.

Talk of impossible worlds as ways things could (absolutely) not have been might suggest that these worlds are, themselves, impossible objects. An impossible object is an object which could not possibly exist and so does not in fact exist. Yet defenders of impossible worlds claim that they do in fact exist. (Or rather, most of their defenders do. Those who don't have a different view of what existence is. We will discuss the issue of the existence of impossible worlds in §2.3.)

This isn't an issue for impossible worlds only. By the same reasoning, we could say: possible worlds (other than the actual world) are merely possible objects and so not actually existing objects. And yet their defenders say that they do in fact actually exist. (Or rather, most of their defenders do. Those who don't have a different view of what actuality is. We will discuss the *genuine realist* approach, on which possible worlds exist but may not actually exist, in §2.2.)

For now, it will suffice to stick to an analogy. Assume that some Escher drawings represent impossible situations. This does not make them impossible. They are not merely possible entities either: they really, actually exist. So actual entities can represent impossible situations. A core part of our investigation into impossible worlds will concern how they manage to represent the situations they represent. (Don't take the Escher drawing analogy too far: we don't want to claim that worlds represent *pictorially*, in the way pictures do.)

The question of how worlds (possible and impossible) represent what they represent is tied up with the question of what they are, metaphysically speaking. This question will occupy Chapters 2 and 3. We'll investigate the issue by first looking at what possible worlds are, metaphysically speaking. Of each of the plausible options, we then ask whether it may be extended to account for impossible worlds.

Some readers may discern no serious issue here. Some modal logicians take the instrumentalist line, on which the set of worlds W may be any old bunch of objects with some relations between them. Anything that does the job will do. This seems philosophically unsatisfactory, just as it seems unsatisfactory to talk of moral properties, or abstract universals, or truth, and yet to refuse to consider their nature. If it is good to understand various concepts in terms of worlds, possible or impossible, then we want to know why this is so. It is difficult to answer the question without saying something about what kind of things worlds are. (Of course, it's often fine to work with worlds without considering their nature, if one is merely postponing, rather than forever refusing to answer, that question.)

Before we get to the logical and philosophical applications of impossible worlds (Parts II and III), therefore, we will investigate the metaphysics of worlds. But, before we get to the metaphysics of worlds and the issue of how worlds represent impossibilities, we should ask whether *we* can represent impossibilities. For if we can't, there is less work to do for impossible worlds in logic and philosophy.

1.5 Conceivability and Possibility

Hyperintensionality is typically thought to involve representational contexts. Impossible worlds have a role to play, first of all, in modelling representational mental states, or thoughts, whose hyperintensional nature is tied to the fact that their content involves absolute impossibilities in some way or other.

But can we actually think about the impossible? Can we have mental representations – intentional states of the mind – directed to impossible contents? A venerable philosophical tradition denies this. Hume is the most quoted authority:

> 'Tis an establish'd maxim in metaphysics, that whatever the mind clearly conceives includes the idea of possible existence, or in other words, that nothing we imagine is absolutely impossible. (Hume 1739/1978, I, ii, 2)

We think that Hume's maxim is wrong (as do Byrne (2007), Fiocco (2007), Kung (2010), and Priest (2016a), among others). Arguing for this requires us to say something about conceivability and imagination. These are highly ambiguous notions. One way to clarify them consists in asking how mental representations in general represent, and looking at answers provided by cognitive psychologists. (We now follow Berto and Schoonen 2018.) The literature presents two main candidate codings for mental representations: the *linguistic* and the *pictorial*, the difference between the two consisting in the degree of arbitrariness of the representation relation (Paivio 1986).

Pictorial mental representations are gathered under the rubric of 'mental imagery', and characterized by reference to sensory perception. They are 'quasi-perceptual experiences' (Thomas 2014, Introduction), for they resemble perceptual representation, but can occur in the absence of the actual stimuli. Studies on neuroimaging, such as Ganis et al. 2004, seem to show that visual mental imagery (the most studied kind of imagery) activates about 90% of the same cerebral areas activated by visual perception, though the interpretation of such results is somewhat controversial.

Visual mental imagery is often claimed to have spatial or quasi-spatial features. When we entertain imagery of this kind, we represent objects and situations typically in three-dimensional egocentric space. These representations are available for 'parallel processing' because they have some kind of mereological structure (Paivio 1986, 198). You can represent to yourself in this way the arrangement of your living room and describe its contents from different viewpoints, mentally

scanning the objects included there from top to bottom or from left to right; you can mentally zoom into a corner, and so on. Of course, psychologists who work on mental images do not claim that they are real pictures, hence the frequent use of the 'quasi-' prefix. The claim that the parts of a pictorial mental representation correspond to the parts of the represented scenario, with the relative distances respected, comes with the proviso that 'part' and 'distance' should be understood functionally rather than spatially (see e.g. Kosslyn and Pomerantz 1977).

Linguistic mental representation, by contrast, is arbitrary in the same way that the connection between words and what they mean is arbitrary. Such representations are called 'amodal' to stress that they are disconnected from sensory modalities in a way pictorial representations are not. According to Paivio (1986, 198), linguistic mental representations are processed serially, the way we process the meanings of sentences through their subsentential components. This is taken as evidence that linguistic representations lack the mereological and quasi-spatial features of (visual) pictorial ones.

Paivio's *dual coding* theory has it that there are precisely two codes for mental representations: the linguistic and the pictorial. Cognition works with two functionally independent (though interacting) systems handling representations of the two kinds. The usefulness of having two systems, according to some, lies in the different contents the two are apt to represent: pictorial imagery is more suitable for concrete situations which are proximal in space and time, whereas linguistic representation works better for abstract scenarios involving non-perceptual features (Amit et al. 2009).

(Some psychologists, including Pylyshyn (1973, 2002), think that there is really just one kind of mental representation. They attempt to reduce the pictorial to the linguistic. This involves the 'imagery debate' or 'analog/propositional debate', to which we return in Chapter 7.)

We argued in Berto and Schoonen (2018) that, if mental representations involved in conceivability represent linguistically, then Hume's maxim cannot even get off the ground. If we make the plausible assumption that linguistic mental representations have at least the

same representational power as the expressions of natural languages like English, then of course we can conceive, by linguistically mentally representing it, the impossible. Logically impossible sentences of ordinary English can be perfectly meaningful.

Quine (1948) argues that contradictions can be meaningful. He makes the point as a response to his fictional philosopher Wyman, sometimes taken as representing Meinong's view (to which we will come back in §2.3) that some things do not exist. Wyman believes that things like Pegasus ought to be admitted in our ontological catalogue, as *possibilia*, for otherwise it would make no sense to say that Pegasus is not. By parity of reasoning, says Quine, we ought to admit the round square cupola on Berkeley College; otherwise, it would make no sense to even say that *it* is not. But accepting this brings inconsistency. Wyman reacts by declaring that inconsistent conditions are meaningless. We find Quine's reply spotless:

> Certainly the doctrine [that contradictions are meaningless] has no intrinsic appeal; and it has led its devotees to such quixotic extremes as that of challenging the method of proof by *reductio ad absurdum* – a challenge in which I sense a *reductio ad absurdum* of the doctrine itself.
>
> Moreover, the doctrine of meaninglessness of contradictions has the severe methodological drawback that it makes it impossible, in principle, ever to devise an effective test of what is meaningful and what is not. It would be forever impossible for us to devise systematic ways of deciding whether a string of signs made sense – even to us individually, let alone other people – or not. For it follows from a discovery in mathematical logic, due to Church (1936), that there can be no generally applicable test of contradictoriness. (Quine 1948, 34–5)

Graham Priest, a friend of true contradictions, agrees (for once!) with Quine:

> If contradictions had no content, there would be nothing to disagree with when someone uttered one, which there (usually) is. Contradictions do, after all, have meaning. If they did not, we could not even understand someone who asserted a

> contradiction, and so evaluate what they say as false (or maybe true). We might not understand what could have brought a person to assert such a thing, but that is a different matter and the same is equally true of someone who, in broad daylight, asserts the clearly meaningful 'It is night'. (Priest 1998, 417)

Now suppose that, instead, there are irreducibly pictorial mental representations. Then the question for supporters of Hume's maxim is: does pictorial imagination work *purely* pictorially, or not? Does the relevant mental imagery represent a situation without any language-like arbitrary assignment of meaning, but just via the phenomenological similarity of the imagery to the worldly situation?

It is controversial whether mental representation can *ever* work purely pictorially. Fodor (1975) argues that pictorial mental representation has a role in cognition only insofar as it works as 'imagery under description', that is, insofar as the imagery comes endowed with linguistic labels: linguistic mental representations pinning down what the image is about. If so, then the arbitrariness of the relevant linguistic labels allows us to imagine the impossible.

Kung (2010) argues that this stipulative labelling component gives pictorial imagination its power to represent the impossible: for example, by stipulating the identity of the imagined objects. Imagine Tim kissing John. The phenomenology of the mental imagery can be such that the represented figures are relevantly similar to Tim and John: hair colour, eyes, bodies. But what makes the imagining count as a representation of a scenario in which *Tim* kisses *John* is that one takes one figure as representing *Tim* and the other as representing *John*. And just as one can imagine Tim kissing John (a possible scenario), so can one imagine Tim as a cleverly disguised robot. One labels the imagined person-lookalike, which turns out to be filled with circuits and transistors, as *Tim*. But Tim (suppose) is essentially human, so this scenario is metaphysically impossible.

What if we do have pictorial mental imagery that represents purely pictorially? Then it may be that scenarios imagined in this way must be possible. But mental imagery of this kind would be quite limited in scope (Berto and Schoonen 2018). Some labelling seems to be needed

whenever perceptual experience has a content that goes beyond mere shapes and colours (Siegel 2006, Siewert 1998). You see a face and a nose, rather than merely face-like and nose-like shapes. Your experience comes labelled: the nose-like shape as a *nose* and the face-like shape as a *face*.

Purely pictorial mental imagery on its own would be limited, in particular, as a tool of modal epistemology. Philosophers discuss whether the imaginability of intrinsic universals, time travel, or a supreme being having all perfections to the highest degree entails their absolute possibility. (Van Inwagen (1998) doubts that we can imagine these things; see Hawke 2011 for a discussion.) Imagination here cannot be purely pictorial, for it involves abstract objects and properties far removed from sensory perception. In these debates, 'imagination' seems to be understood more broadly than the purely pictorial characterization allows.

If, finally, mental representations are taken to be neither linguistic nor pictorial, this leaves the supporters of Hume's maxim with a heavy burden of proof. They seem forced to invoke a peculiar 'third code' of representation, with no counterpart in general theories of representation. It's then up to Humeans to provide a plausible theory of how that notion works. Absent a workable theory, the approach is relying on representational magic. In short, we have good reason to hold that we can mentally represent absolute impossibilities.

Chapter Summary

Possible worlds are ways things might have been (§1.1). They find applications in analysing possibility and necessity; propositions; knowledge and belief; information; and indicative and counterfactual conditionals (§1.2). But possible worlds semantics faces the issue of hyperintensionality, generated by concepts that require distinctions between logical or necessary equivalents. The problems of distinguishing equivalent propositions, of logical omniscience, of information overload, of irrelevant conditionals, and of counterpossible

conditionals, are all instances of the general issue (§1.3). Adding impossible worlds promises to help with these puzzles (§1.4). But can we genuinely think about the impossible? We argued that we can (§1.5).

2

Metaphysics

2.1 Ways of Thinking about Worlds

We now investigate the metaphysics of possible and impossible worlds. What are such worlds like? And how do they manage to represent whatever it is they represent? (A difficult question concerning the *granularity* with which impossible worlds represent what they represent will appear in §8.4.) Our general approach will be to begin with a theory of possible worlds, and to ask whether it may be extended to accommodate impossible worlds. We will structure our discussion around four central questions:

(Realism) Should we be realist or anti-realist about non-actual worlds?

(Existence) Do non-actual worlds exist, or are they non-existent entities?

(Genuineness) Are non-actual worlds genuine worlds or ersatz entities?

(Parity) Should we give the same answer to the three questions above for both possible and impossible worlds?

While the first three questions require answers from any possible worlds theorist, the fourth one is specifically addressed to impossible worlds theorists like ourselves.

The first two questions – REALISM and EXISTENCE – sound similar, but they might differ in an important way. REALISM asks whether we should talk about non-actual worlds when we're being most metaphysically serious: should we quantify over non-actual worlds at all, using our most serious metaphysical quantifiers, and therefore (following Quine (1948)) accept commitment to such things in our metaphysics? If we answer *yes* to REALISM, then EXISTENCE asks, should we say those worlds exist, or should we instead say they are non-existent entities? Now according to Quine himself, and many others (e.g. Van Inwagen (2008)), one who answers *yes* to the first question is automatically committed to the existence of the relevant objects, for existence is captured by the quantifier. But, as we will see in 2.3, others disagree on this reduction of existence to quantification.

Let's probe the question of REALISM a little further. People draw the realism/anti-realism contrast in different ways. Sometimes, anti-realism about X simply means the view that the X's aren't objects of genuine, ontologically committing quantification. Other times, it means that the X's are mind-dependent parts of reality. (This is what 'anti-realism' often means in logic and mathematics. The view that numbers exist, but are constructions of the human mind, is often called anti-realist.) Here, we'll use 'anti-realism' about worlds in the former sense.

All participants in our discussion agree that we get to indulge in worlds-talk. *Realists* interpret that talk literally: we're talking about parts of reality, which we can refer to and quantify over when we're being at our most metaphysically serious. Anti-realists, by contrast, think that we get to indulge in worlds-talk but without ontological commitment to non-actual worlds.

Let's now turn to the remaining two questions, GENUINENESS and PARITY. GENUINENESS asks: should we understand the non-actual worlds as being on a par, metaphysically speaking, with our own world? Or should we accord them some other metaphysical status? Exactly what this means is a delicate matter. We'll discuss it in §2.2, where we also explain the terminology of 'ersatz' and 'genuine' worlds. PARITY is a more straightforward question: should we treat impossible

worlds as being metaphysically on a par with the merely possible ones, so that the answers to the first three questions are the same for both kinds of worlds? Or do impossible worlds require their own special metaphysical treatment?

We won't consider all 16 combinations of answers to these questions. Some combinations do not make much sense, and some haven't been discussed at all in the literature. (Aside to potential PhD students: there are unexplored possibilities here!) We'll focus on those combinations which have received the most attention.

2.2 Genuine Realism

Genuine realism says that there exist non-actual worlds very much like our own. The view usually covers just the possible worlds; applying it also to impossible worlds gives *extended* genuine realism. We'll first explain what the 'genuine' bit in 'genuine realism' is supposed to mean. After that, we'll take a look at some of the theories falling into this camp.

The (unextended) view is often put by saying that these worlds are *concrete* entities. This won't quite do for the extended version, but it nevertheless provides a good conceptual entry point to the idea. On this approach, it's possible that there's a talking wombat because there exists a possible world which has a talking wombat as a part. That wombat is a real, flesh-and-blood living creature, just like our wombats. It's located in space and time, and is part of the causal order of that world. That's the sense in which that wombat is a concrete entity, and that's also the sense in which worlds made up of such entities are concrete worlds.

(This is a good time to Google-image-search 'wombat', if you haven't done so already. We'll meet you back here in a bit.)

However, it would be a mistake to identify the genuine worlds with those made up of concrete entities. When we turn to impossible worlds, we will need to make room for worlds which differ on mathematical and logical facts. Such facts, and the entities they

concern, are typically viewed by philosophers as being non-causal and located outside time and space. We might, for example, have cause to consider the impossibility that 3,456 is the largest natural number. The genuine realist won't analyse this impossibility in terms of a concrete world which has the number 3,456 as a concrete part. The intuition is: even if 3,456 were the largest natural number, it would still be an abstract mathematical entity. So we shouldn't identify the genuine worlds as those that are concrete. (Note that this problem doesn't arise with the unextended view. Possible worlds never differ over logical or mathematical facts, and so, on the unextended view, mathematical and logical entities don't need to be treated as parts of worlds at all.)

For a better analysis of 'genuine', let's return to our talking wombat world. It's genuine just in case it contains a real talking wombat. It represents the existence of a talking wombat by having a real talking wombat as a part. Similarly, a genuine world represents that 3,456 is the largest natural number by having 3,456, but no larger natural number, as a part. GENUINENESS is a matter of how a world represents, not of whether it is a concrete or an abstract entity. Worlds that do not represent genuinely are *ersatz* worlds. The key distinction between genuine and ersatz (non-genuine) worlds concerns how those worlds represent. Genuine worlds represent (*de dicto*) possibilities and impossibilities directly, by having them as parts. Non-genuine ersatz worlds represent them in some other way.

Now let's take a look at some of the theories falling under the banner of genuine realism. We'll begin with unextended views covering just the possible worlds, and then see whether they can be extended to cover impossible worlds.

Lewisian Realism

Lewis (1986b) views possible worlds as wholes, each unconnected in space and time from the others, whose parts are themselves concrete entities located in space and time. Take any spatiotemporal entity, and take all those entities related to it in space and time. All of those taken together make for a possible world. So for each world, every

part of it is spatiotemporally related to every other part of it, but not to any part of any other world.

The answer to our GENUINENESS question, on this view, is that non-actual worlds differ from the actual one only in terms of what goes on at them, not in kind. Indeed, the actual world has no ontological privilege in this modal metaphysics. According to Lewis, 'actual' works like an indexical expression. Just like 'here' and 'now' refer to the place and time of utterance, so does 'actual' refer to the world of utterance. The possibility of there being talking wombats is represented by some non-actual world including talking wombats as parts. From our own viewpoint, those chatty wombats are non-actual *possibilia*: things that lack actual existence. But from their viewpoint, they are the actual folks, and we are non-actual *possibilia*.

This approach gives Lewis a reductive definition of *possible world*. According to him, one main advantage of genuine modal realism is in providing an extensional, non-modal account of modal notions. Take a (non-world-indexed) definition of absolute (*de dicto*) possibility as unrestricted quantification over possible worlds:

(P) $\Diamond A$ is true iff A is true at some world w

Whether this equivalence provides a reduction of possibility to non-modal concepts depends on whether the notion of *world*, involved in the quantification on the right hand side, is itself modal. Lewisian possible worlds – maximal mereological sums of concrete, spatiotemporally connected entities – are wholly extensional. And if certain Lewisian criticisms of ersatzism from his (1986b) are right, then each ersatz account of worlds on the market may have to resort to primitive intensional entities (such as propositions) and to primitive modal notions. We will return to this issue in Chapter 3, where we investigate ersatz theories in detail.

It is controversial whether Lewisian modal realism *can* indeed provide an accurate, reductive, non-modal account of modal notions (Divers and Melia 2002). Even if it can, this advantage is put at stake by adding genuine impossible worlds. Once such worlds enter the stage, (P) becomes false from right to left, insofar as the quantifier

on the right ranges over all worlds. One then needs a principled way to restrict that quantifier to possible worlds. Achieving this without resorting to primitive modal notions can be a tricky issue. The Lewisian reductive definition of *possible world* won't work for impossible worlds, because (as discussed above) these need to include non-spatiotemporal entities, such as mathematical entities. So, in extending the account to include impossible worlds, one may have to give up on this reductive ambition. The view would then say that worlds are wholes, with both spatiotemporal and non-spatiotemporal entities as parts.

It's not even clear that we can assume that the spatiotemporal portion of a world will include any entity spatiotemporally connected to any other part of that world. Such worlds are *spatiotemporally maximal*. It seems impossible for reality not to be spatiotemporally maximal. But then, there will be genuine impossible worlds which aren't spatiotemporally maximal.

There is an even more serious worry for the extended Lewisian approach, however. Lewisian worlds (possible or impossible) obey:

(EXPORTATION) If world w represents something as being F, then something is F (simpliciter).

To see why this is so, suppose Lewisian world w represents something as being F. Then, by definition, w contains an F as a part. As w is part of reality, by the (fairly uncontroversial) transitivity of parthood, that F too is part of reality. It exists, full stop, in just the same sense that the city of Nottingham exists. Because Lewisian worlds are genuine worlds, representing possibilities directly by having them as parts, the represented objects are really out there.

The exportation principle is problematic for any account of impossible worlds, as Lewis remarks in a famous footnote (Lewis 1986b, 7, fn 3). Exporting merely possible entities or states of affairs from genuine possible worlds lumbers us with a large and counterintuitive, but still consistent ontology. (Let's bracket any inconsistency worries raised by the specifics of Lewis's theory. See

Jago 2016 for discussion.) Exporting impossible entities or states of affairs from genuine impossible worlds, by contrast, drags us into contradiction. There is an impossible world at which there is a round square. If that world is genuine and we can export from genuine worlds, then there really is an entity which is both round and square. But it is a necessary truth that no square is round (that's why the round square was impossible to begin with!), and so our exported round entity is also not round. Consequently, it both is and is not round: contradiction.

Generalizing on this point, 'at world *w*' will distribute across conjunction and negation when *w* is a genuine world. That's because a genuine world *w* represents something as being $F \wedge G$ by containing a part that's both *F* and *G*; and it represents something as being *F* iff it has a part that's *F*. So, in particular, to represent something as being such that $Fx \wedge \neg Fx$, *w* must contain a part that both is and is not *F*. World *w* represents that thing as being *F*. But since that thing is also not *F*, it's false that world *w* represents it as being *F*. That's a straightforward contradiction: 'at *w*, $Fx \wedge \neg$(at *w*, *Fx*)'. Lewis presents this argument as part of his reason for having no use for genuine impossible worlds in the first place. It will not help to claim that one can speak truly by contradicting oneself (only) when the subject matter is an impossibility (Yagisawa 1988): we are still committed to there being true contradictions, which is an unwelcome conclusion to anyone who is not a dialetheist, à la Priest (1987).

Yagisawaian Realism

Yagisawa (2010) gives an alternative to the Lewisian view of worlds. His account is particularly interesting for our purposes, because he focuses explicitly on impossible worlds. Yagisawa treats modality much as *four-dimensionalists* treat temporal matters. On the latter view, entities exist and have properties at a time *t* by having *temporal stages* at time *t* which have those properties. Lenny is Schnauzer-shaped this Monday in virtue of having a Schnauzer-shaped this-Monday-stage; he was once a puppy in virtue of having a past

puppy-stage; and he is always adorable in virtue of all his temporal stages being adorable.

Similarly, says Yagisawa, entities exist and have properties at a world *w* by having *modal stages* at world *w* which have those properties. Lenny is actually Schnauzer-shaped in virtue of having a Schnauzer-shaped actual-stage; he could have been portly in virtue of having a (merely) possible portly world-stage; and he is necessarily canine because all of his world-stages are canine.

We're unsure whether this approach avoids the exportation worry. (The following draws on Jago 2013a.) In the temporal case, some temporally extended entity such as Lenny has properties-at-time-*t* in virtue of having *t*-stages with those properties. This reduces Lenny's properties-at-a-time to properties had by his temporal stages. His Monday-stage is Schnauzer-shaped, simpliciter; that stage is intrinsically Schnauzer-shaped.

Similarly, in the modal case, Lenny has properties-at-world-*w* in virtue of having *w*-stages with those properties. His *w*-stage is portly, simpliciter; that stage is intrinsically portly, even though Lenny (the collection of all his stages) is not. So the possibility of Lenny's being a portly Schnauzer entails that there is a portly Schnauzer stage, perhaps not actually, but out there somewhere in modal space. That stage is intrinsically portly and not merely portly-at-*w* (for some world *w* or other). But by the same token, the impossibility of Lenny's being a portly-and-slim Schnauzer entails that there is an intrinsically portly-and-slim Schnauzer stage, certainly not actually, but out there somewhere in modal space. We can truly say that that stage of Lenny is both portly and not portly: we have not avoided contradiction.

Yagisawa (2015) responds to this objection on the grounds that

> It is integral to [extended genuine realism] that predications in modal metaphysics be made with careful attention to … modal tense … on verbs in discourse concerning goings-on at worlds and goings-on in modal space at large. (Yagisawa 2015, 319)

He describes four modal tenses: 'actuality tense, mere-possibility tense, (metaphysical) impossibility tense, and a modal tense specifically

for predications concerning modal space at large' (Yagisawa 2015, 319). The idea is that facts about Lenny's contradictory impossible-world stage are expressed using the impossibility tense, from which (according to Yagisawa) we cannot infer contradictions simpliciter.

This is a complex proposal, and we shan't evaluate it in detail. We'll merely note that whether Yagisawa's approach really does avoid the worry depends on what we are allowed to express in the 'modal space at large' tense (which Yagisawa marks with an '*m*' subscript). This is the tense in which we express what modal space at large looks like: it contains possible and impossible worlds, with modal stages of individuals at those worlds. Such facts are expressed in the *m*-tense. Since we express the existence of those worlds and the modal stages at them in the *m*-tense, it is natural also to express facts about those worlds and states in that tense.

If so, we'll be able to express the existence of the Lenny-stage that both is and is not portly in that tense. But then, Yagisawa's story about modal space is itself contradictory, which is precisely what he wants to avoid. If, however, facts about Lenny's impossible world-stages may be expressed only in the impossibility-tense, then this argument will not go through. It is clearly no objection to the account that impossible words are contradictory!

McDanielian Realism

One way to avoid EXPORTATION is to deny that entities like Lenny have this or that property simpliciter. Rather, on this approach, they will have some properties relative to one world, but different properties relative to other worlds. So, rather than asking what Lenny is like in and of himself, we will need to ask what he is like at this or that specific world. So we can say that Lenny is portly relative to world w, but not portly relative to w_1. He then has the relational properties, *being portly-at-w* and *being not-portly-at-*w_1. These properties are not in conflict with one another (since $w \neq w_1$), and so no contradiction arises. McDaniel (2004) defends a view along these lines, sometimes called *modal realism with overlap*. (Transposed to the temporal case, this is similar to the *three-dimensionalist* view.)

Although McDaniel focuses only on possible worlds, the view is interesting from an impossible worlds perspective, because it blocks EXPORTATION. If Lenny is both portly and not portly at an impossible world *w*, then he has the property of *being both portly and not portly at world w*. But we can't infer from this that Lenny is and is not some specific way. It does not follow, for example, that he both possesses and does not possess *being portly-at-w*. World *w* is an impossible world, remember, and so there is no guarantee that *being not-F-at-w* amounts to lacking the property *being F-at-w*. So the argument to contradiction we discussed above is blocked, on this view.

There is a problem, however, in attempting to extend McDaniel's view to incorporate impossible worlds. (Here we draw on Jago 2014a.) It cannot be the case that Iggy Pop is a singer but James Newell Osterberg isn't (since Osterberg is Iggy Pop). That's impossible. So there is an impossible world *w* according to which Iggy Pop but not Osterberg is a singer.

On the current analysis, that is to say that Iggy Pop, but not Osterberg, bears the *being a singer* relation to world *w*. Then Iggy Pop bears a relation to *w* which Osterberg does not bear to *w* and hence, by Leibniz's law, Iggy Pop and Osterberg are not identical, simpliciter. (Note that our use of Leibniz's Law here is in an extensional context, in which we consider relations between an individual and a world.) But this is absurd, for Osterberg *is* Iggy Pop! On that basis, it seems difficult to extend McDaniel's account to include impossible worlds.

Lewisian, Yagisawaian, and McDanielian realism all have trouble with impossible worlds, although the problems for Yagisawa's view may not be insoluble. Perhaps this suggests that we should treat possible and impossible worlds differently. We consider that option in §2.5. Before that, we'll consider alternative answers to EXISTENCE and GENUINENESS.

2.3 Non-existent Worlds

In this section we use material from Berto and Plebani 2015, chapter 7. As we claimed above, the question of REALISM may or may not be distinct from the question of EXISTENCE, depending on one's view of what existence is (or of how existence claims are best expressed). If one takes the notion of existence to be best expressed by quantified statements, as Quine (1948) did, then genuine uses of quantification over non-actual worlds will involve commitment to the existence of non-actual worlds.

One could, however, subscribe to a different notion of existence. One could take existence to be a *real property* in the Kantian sense: a genuine feature that some things have, but others lack. If some things do not exist, then existence will not be captured by the quantifier. In this way, one can be a realist about non-actual worlds without automatically committing to their existence. In this setting, the questions of REALISM and EXISTENCE come apart.

The view that some things do not exist is nowadays often labelled as 'Meinongianism', after Meinong (1904). (See Jacquette 1996, Parsons 1980, Routley 1980, Zalta 1983) Meinongians distinguish the *Sein* of objects – their existential status – from their *Sosein* – their having features or properties. Things can bear properties even when their existential status is *none*, when they lack the feature of existing. The view is sometimes interpreted as making a distinction between *being* and *existence*. In this interpretation, it is granted that whatever is quantified over in literally true sentences must have some being (after all, it is claimed that there *are* things which are such-and-such), though it may lack that more accomplished way of being we ordinarily call existence.

However, several Meinongians, including Berto (2012) and Priest (2016b), deny any such distinction between being and existing: they claim that some things have no being-or-existence whatsoever (and if there are different ways of being or of existing, some things have none of them). As for the verb 'to be' showing up in 'there are', they claim that it is accidental to quantification: its appearing in some of

the quantificational expressions we use in ordinary language lends thin linguistic support to the thick ontological view that whatever we quantify over must have being.

English also uses 'some', where the verb 'to be' does not appear. German often uses *es gibt*, but we would hardly infer that Germans ascribe *giving*, or *being given*, to what they quantify over. French often uses *il y a*, which includes the other auxiliary verb, 'to have', but we would hardly infer that the French ascribe *having* to what they quantify over. Besides, we use 'there is' in locative constructions where 'exists' make no sense, which casts doubts on their synonymy. Compare, 'there was a girl in the office this morning; while she was there, she was looking for you' with 'a girl existed in the office this morning; while she was existing there, she was looking for you'.

What kinds of things can lack existence? The most straightforward candidates have traditionally been fictional objects: things described in tales, stories, fantasy novels, like Sherlock Holmes, Heathcliff, Gandalf, and Phlebas the Phoenician (we will come back to them in Chapter 11). Other candidates come from temporal and modal considerations: past existents like Heraclitus (he does not currently exist, though he does still bear features, like being Heraclitus, or being admired, or being obscure and often misinterpreted); future existents like the first newborn of the twenty-second century; or merely possible objects lacking actual existence, like Wittgenstein's daughter (Wittgenstein died childless, but he could have had a daughter), or the eight-legged dog statue that could have been made from the marble out of which Michelangelo actually sculpted David.

Parsons (1980) and Priest (2005) propose that worlds be understood as nonexistent objects. Of all the worlds, just one, these physical surroundings of ours, has the feature of existing. All the others lack it. One way to mark the difference between what is actual and what is not, while leaving room for the non-actual, is to identify the actual with the existent. One can then claim that existence (and so actuality) is not all there is. Non-actual worlds are nonexistent objects which are, in some sense, maximal. Priest's (2005) Meinongian account comes endowed with both possible and impossible worlds.

Some worry that nonexistents (including nonexistent worlds) have no clear identity conditions, and that we cannot know about such entities as they are devoid of spacetime location and causal powers. But these difficulties also apply to other realist accounts of worlds. Usually ersatz accounts (§2.4 and Chapter 3) have it that worlds are abstract objects. If they are constructions out of propositions or maximal property-like entities, then a rigid Quinean may ask for plausible identity criteria for things of these kinds before accepting ersatz worlds. (For an extended discussion of identity criteria for nonexistent objects, see Berto 2012, part II and chapter 8.) Abstract objects are also devoid of causal powers and spatial location. On the other hand, Lewisian genuine possible worlds are concrete, but by definition causally and spatiotemporally isolated from one another and from us. So epistemic access to worlds is problematic whether they are Lewisian, abstract, or nonexistent. Overall, it is not clear that a Meinongian view of worlds as nonexistents is worse off than other realist accounts.

Priest argues that modal facts can be known, on the Meinongian view of worlds, largely by stipulation and imaginative exercise. This, he claims, is similar to how we know things about nonexistent fictional characters. Conan Doyle was free to stipulate that Holmes lived in Baker Street, rather than Oxford Street (Priest 2008, 31). Similarly, we can stipulate worlds that are particular ways: for instance, such that Trump lost the US presidential election. Just as we can stipulate that Holmes lived in Baker Street, so can we stipulate that Hobbes squared the circle. But a world where one can square the circle with ruler and compass is not a possible world. So stipulation can give us nonexistent worlds that represent possibilities and impossibilities. How do we know which is which? That is a difficult question of modal epistemology, but it might be that answering it is no more problematic than for other views.

What is harder for the Meinongian view, we submit, is to give a precise explanation of how worlds represent (or realize) possibilities and impossibilities. The Meinongian view complicates the issues surrounding GENUINENESS. In a sense, there *is* a metaphysical (or,

ontological) difference between the actual world, which exists, and the rest, which do not. But given how Meinongians understand existence, this does not tell us anything about how nonexistent worlds represent what they represent.

It is difficult to defend the view that nonexistent worlds can represent as genuine worlds, that is, by realizing the relevant possibility or impossibility as a part. Some problems for this view look similar to the ones examined above for genuine realism, having to do with the EXPORTATION principle. A genuine nonexistent impossible world would represent there being round squares by having real round squares as parts. Those round squares are nonexistent, but they really are both round and square. So the Meinongian who takes this stance is committed to move from asserting 'world *w* represents something as being round and square' to asserting 'something is both round and square'. In other words, she is committed to EXPORTATION, just as Lewisians are, and so it stuck with true contradictions. (This may be no problem for dialetheists such as Priest (1987), but it is for anyone who aims for a consistent metaphysics of worlds.)

One may ask what it means for a nonexistent entity to have parts. The idea of parthood is most intuitive when it involves concrete existents, such as this table having its legs as parts. One could gesture to an answer by claiming that parthood relations make sense for abstract objects (as in, 'the antecedent is part of the conditional'), which also lack spatial features. But it is an open issue to what extent this is more than a metaphor. (What does 'part' really mean when we say that writing accurate reviews is part of being a good referee?)

Another worry is that we want worlds to represent things as existing. For a genuine world to represent something as existing – Wittgenstein's sister, say – is to have a sister of Wittgenstein's as an existing part. But how can a wholly nonexistent world have a part which exists? Moreover, given EXPORTATION for genuine Meinongian worlds, that sister of Wittgenstein exists, simpliciter. But this undermines talk of taking mere possibilia to be nonexistent entities. One may think that we should not infer from the existence of the parts to existence of the whole. In general, it's fallacious to infer a

property *F* of the whole from its parts being *F*s. (Thanks to Graham Priest for suggesting this move.) However, a whole with an existent part is surely something that *in part* exists, whereas a Meinongian should say that her worlds are wholly nonexistent entities.

Given these worries, one may think that nonexistent worlds had better represent as ersatz worlds, without realizing the relevant possibilities or impossibilities. (We'll discuss ersatz approaches in §2.4 and then again in Chapter 3.) Nonexistents are often invoked as representational targets: we seem to be able to speak about them, quantify over them, and also to intend them in our thoughts (Crane 2013). Nonexistents are introduced as what is represented, rather than as what does the representing. An associated account of representation is lacking. This is not a refutation of such Meinongian approaches to worlds as nonexistents, but a challenge to be addressed.

2.4 Ersatz Modal Realism

On our understanding of what it takes for a world to be genuine, it must represent the existence of an *F* by having an *F* as a part. Such worlds represent the existence of a talking wombat by having a real talking wombat as a part. Ersatz worlds, by contrast, represent the existence of an *F* in some other way. We are using 'ersatz world' as a catch-all term, to cover worlds which represent such-and-such, not in the way genuine worlds do. To get a sense of how this might go, consider how a story represents some event. It does so by being composed of sentences of some language, whose meaning conveys that such-and-such happened. It's clear that the story can exist even if the things and events it describes do not, and never have: that's what it is to be a fiction. We have no ontological trouble with stories of hobbits, insofar as we can make sense of them without thereby committing ourselves to the existence of hobbits.

We can think of worlds on that model. They exist, and they represent such-and-such as existing and as being certain ways. But they don't represent those things being those ways by having those

things being those ways as parts. So we should feel no urge to infer from the existence of those representations, of hobbits, say, to the existence of hobbits.

In general, when the relevant world w is ersatz, EXPORTATION will not hold. This helps especially with impossible worlds: w can consistently represent inconsistencies, such as *that x is both F and not-F*, without implying the reality of any such x. This approach, *ersatz modal realism*, is thus compatible with *actualism* in metaphysics: the view that nothing exists but what actually exists. In particular, ersatzists say that their worlds, and all their constituents, actually exist and are part of the actual reality surrounding us. (Again, compare the situation with stories. Their constituents are the words that make them up, all of which actually exist.)

How can our world include a plurality of worlds within itself? There seems to be insufficient room in actuality for that! The standard ersatzist reply (Divers 2002, chapter 10) consists in distinguishing between (a) actuality, this reality surrounding us and which, for actualists, is everything there is; and (b) what is *actualized*. Of the various representations of how things could have been, one stands out as representing actuality precisely as it is, and this is the one which is actualized.

Ersatz modal realism can then analyse possibility and necessity in the standard way, as existential and universal quantification over possible worlds, respectively (§1.2). Ersatzists typically take their worlds to be abstract entities: maximal properties, or sets of sentences or propositions. (We will get to the details in Chapter 3.) In this way, ersatzists typically have the resources to include impossible as well as possible worlds in their ontological toolkit, for example, as sets of 'worldmaking' sentences which include both A and $\neg A$. This encourages a positive reply to the PARITY question. Once one has the relevant abstract objects at one's disposal, one can put them to work in the representation of impossibilites as well as possibilities, at no extra ontological cost. (Just as one can put a storytelling language to work to compose stories that speak of impossible as well as possible happenings.)

Now go back to the project of giving a reduction of modal notions to purely extensional ones. Taking again our clause for unrestricted possibility:

(P) $\Diamond A$ is true iff A is true at some world w

How should we delimit the quantification on the right-hand side, so that it ranges only over ersatz *possible* worlds? This seems hard to achieve if worlds are constructions out of abstract entities such as properties, propositions, or sentences. A widespread answer among ersatzists consists in biting the bullet (Vander Laan 1997). This approach accepts that no complete and accurate reduction of modal notions to non-modal ones is feasible. Nevertheless, say the defenders of this approach, conceptual elucidation may come from an analysis given in other, allegedly better understood, notions – even if they are modal concepts. This answer can be paired with a *tu quoque* argument, showing that genuine realism too cannot achieve a fully extensional modal reduction (Divers and Melia 2002).

The big question for ersatzists is: *how* do ersatz worlds represent such-and-such as being the case? Different answers give us different versions of ersatz modal realism. Here are some ways to represent *that A*:

(STATE) By using a state of affairs of a certain kind.

(PROPERTY) By using a property reality would have, were things such that *A*;

(COMBINATORIAL) By taking objects and properties which, if recombined in a given way, would make it the case that *A*;

(MAP) By using a map, picture, or image which depicts things being such that *A*;

(PROPOSITIONAL) By using the proposition or some other content-carrying entity *that A*;

(LINGUISTIC) By using bits of language, whose meaning is *that A*;

(PRIMITIVE) By taking the relevant representation to be a basic, unanalysable feature.

Each of these ways of representing that *A* gives us a form of ersatz modal realism: we'll call them *property ersatzism*, *combinatorial ersatzism*, *map ersatzism*, and so on. Ersatz modal realism, in its various guises, provides a rich resource for theorizing about possibilities and impossibilities. Because of the range of options available, we're going to postpone discussion of the various ersatz accounts until Chapter 3, which will be wholly dedicated to evaluating the approaches in detail.

We're unsure whether the final option just mentioned, *primitive ersatzism*, should be counted as a form of ersatz modal realism at all. To be sure, it's a realist view without genuine worlds: in that sense, it's a form of ersatzism. The source of our hesitation is that it refuses to answer the central question for ersatzists, 'how do worlds represent?' We'll treat this option as a separate account, in §2.7 below. Before that, we'll quickly examine a metaphysics of worlds which is intermediate between the genuine and ersatz approaches.

2.5 The Hybrid View

Hybrid modal realism endorses a hybrid view of what possible and impossible worlds are and of how they represent. It gives a negative answer to our fourth metaphysical question, PARITY, and denies what Priest (1997b) and Rescher and Brandom (1980) call the *Parity Thesis*. Possible and impossible worlds are not on a par metaphysically: the answer to at least one of our previous three questions is different, depending on whether we're talking about possible or impossible worlds. The view is defended in Berto 2010, following suggestions from Divers (2002, Chapter 5) and Kiourti (2010, Chapter 3).

According to hybrid modal realism, both possible and impossible worlds are real and both exist. (That answers REALISM and EXISTENCE.) However, they represent in different ways (GENUINENESS).

Possible worlds are taken as concrete, genuine Lewisian worlds with their real inhabitants, the Lewisian *possibilia*. Impossibilities are represented by set-theoretic constructions from genuine worlds. Genuine worlds represent possibilities by realizing them, whereas impossible worlds are ersatz set-theoretic constructions. They represent impossibilities without realizing them. All possibilities really exist out there in some disconnected spacetime. But there are no real impossibilities, only set-theoretic constructions which merely represent ways things could not have been.

Berto (2010) shows that the hybrid account can distinguish between certain impossible propositions, such as *that swans are black and not black* and *that John is a married bachelor*. The former is a partition on possible worlds, into those where swans are black and those where swans are not black. The latter is a partition on possible worlds, into those where John is a bachelor and those where he is married. Since these are distinct partitions, we have distinct propositions.

The hybrid account also avoids resorting to primitive modality (at least to the extent that Lewisian modal realism does). It is a fully extensional ontology of genuine Lewisian worlds and sets. In our clause for possibility (P), $\Diamond A$ is true iff A is true at some world w, we take the quantification on the right-hand side to range over the genuine Lewisian worlds only. Since these are characterized extensionally (as maximal mereological sums of spatiotemporally related *concreta*), this restriction of the quantifier doesn't resort to any primitive modal notions.

Like other accounts which use ersatz impossible worlds, hybrid modal realism has no problem with the EXPORTATION principle. Inconsistencies and impossibilities at impossible worlds do not spill over into the actual world, nor into any genuine possible world. Impossible worlds are world-stories: abstract set-theoretic constructions. 'At impossible world w, A' means 'according to ersatz-world-story w, A'. From the fact that, according to ersatz world w, A and not-A, it does not follow that: according to w, A and it is not the case that according to w, A.

Hybrid modal realism has other problems, though. One is that it asks us to buy the Lewisian ontology of infinitely many concrete disconnected spacetimes and *possibilia* inhabiting them. Few have accepted such an ontology. Perhaps the most resilient attitude towards Lewisian modal realism is what Lewis himself called 'the incredulous stare'. In spite of Lewis's (1986b) rebuttal of several objections, philosophers remain unconvinced.

Another problem is the extent to which the account can be developed, in order to make all the hyperintensional distinctions we may want to make. (Here we draw on Jago 2014a, 110–12.) In its current form, hybrid modal realism can discriminate between absolute impossibilities on the basis of logical propositional structure. But what about non-logical impossibilities? Hybrid modal realism cannot distinguish between the propositions *that Hesperus is the second planet from the Sun* and *that Phosphorus is the second planet from the Sun*. Since (necessarily) Hesperus is Phosphorus, there is but one planet from which to construct ersatz impossible worlds. So no impossible world contains Hesperus but not Phosphorus, and no impossible world says that Hesperus is not Phosphorus. Yet we want impossible worlds which do distinguish Hesperus from Phosphorus. After all, people believed they were distinct. One may believe herself to be looking at Hesperus but not Phosphorus (or perhaps, she has no particular attitude towards Phosphorus at all).

These worries apply to the specific version of hybrid modal realism presented in Berto 2010. But the idea behind the approach is much more general (see Reinert 2018 for an insightful discussion): start with Lewisian possible worlds, and from them construct ersatz impossible worlds. Different ways of achieving the construction give different versions of hybrid modal realism. If any actualist ersatz approach can be made to work, then hybrid modal realism can too, since it has all the actualist's ontological resources (and then some).

Perhaps what's really at stake in the debate over hybrid modal realism is the issue of modal reduction. This is the hybrid modal realist's key advantage over actualist ersatz accounts. If it turns out that actualist ersatz accounts can define 'possible world' using

only non-modal concepts, or if it turns out that so doing isn't that important, then hybrid modal realism loses its trump card. But otherwise, the view is appealing.

2.6 Encoding Worlds

We've discussed both genuine and ersatz accounts of worlds, distinguished by asking whether worlds represent, say, the existence of a talking wombat by containing a talking wombat. Genuine worlds that represent (*de dicto*) *that A* are such that *A*; whereas ersatz worlds may represent *that A* without being such that *A*.

In Zalta's powerful *abstract object theory* (Zalta 1983, 1997), however, 'being such that' is ambiguous. More generally, saying that some object *o* is *F* is ambiguous. This approach is interesting for our purposes, for it offers a conception of possible and impossible worlds which promises to capture some of the advantages of genuine worlds, but without the pitfalls we discuss in §2.2. (Berto and Plebani (2015, chapter 7) give an introduction to Zalta's view.)

Zalta's central claim is that '*o* is *F*' is ambiguous, between (i) *o*'s *exemplifying* (or *possessing*, or *instantiating*) property *F*; and (ii) *o*'s *encoding F*. Encoding is a primitive notion, and applies only to abstract (non-spatiotemporal and non-mental) objects. For an abstract object to encode a property means, roughly, that that object is partly defined in terms of, or determined by, that property. Abstract objects both exemplify and encode properties, and may encode properties they do not exemplify and vice versa.

Within this theory, *situations* are defined as abstract objects that encode states of affairs, taken in their turn as 0-ary properties, of the form *being such that* —. Worlds are maximal situations (encoding, for each *A*, either the state of affairs *that A* or *that* ¬*A*). Those that could obtain are the possible worlds, and those that remain are the impossible worlds.

This characterization of worlds bears some similarity to (our version of) Plantinga's approach, which we'll discuss in §3.2. On

Plantinga's approach, worlds are identified maximal states of affairs, whereas on Zalta's, worlds are abstract entities which maximally encode states of affairs. So much of what we'll say about the features of Plantinga's approach in §3.2 applies to Zalta's approach as well.

Zalta (1997) is keen to emphasize that his is not an ersatz conception of worlds (even though they are abstract entities). A given state of affairs, *that A*, obtains at world *w* (possible, or not) when *w* encodes the property *being such that A*. And something is such that *A* (at *w*) just in case the state of affairs *that A* obtains (at *w*). In the encoding sense of 'is', a world *w* which encodes *being such that A* is indeed such that *A*. So, on the 'encoding' disambiguation of 'is', Zalta's worlds are genuine worlds. But on the 'exemplification' disambiguation of 'is', they are not.

To illustrate the point: all possible worlds exemplify *being abstract*, but may encode *being concrete*. None exemplify *being such that there is a talking wombat*, else we could infer the existence of that clever marsupial. But some encode this property, for there could have been a talking wombat.

With the notion of encoding, Zalta's theory has a unique take on how worlds represent possibilities and impossibilities. Unlike the genuine worlds discussed above, Zalta's talking-wombat-world does not have a talking wombat as a part. Abstract objects do not have concrete things as parts. And, unlike many of the ersatz worlds we'll discuss in Chapter 3, Zalta's worlds are not constructed by representational entities.

Zalta's object theory is designed to deliver a wide range of benefits, including a theory of mathematical entities. The account of worlds is something of a by-product of the overall theory. To evaluate the approach properly, we'd need to assess its benefits across those areas. In particular, we'd need to assess whether those benefits justify treating 'is' as ambiguous. But that's well beyond the scope of our discussion here. We'll merely note that positing an ambiguity, with little or no linguistic evidence, is always hard to justify.

To highlight the worry, consider some reasoning involving 'is': Lenny is barking; therefore, something is barking. That seems a priori

valid. But if it is, then the 'is' here cannot have the 'encoding' reading, for that reading would not support the inference. (Recall that some worlds encode *being such that there is a talking wombat*, but we cannot infer that something is such that there is a talking wombat.) And if an ordinary use of 'is' like this one cannot have the 'encoding' reading, then positing a general ambiguity looks a doubtful move.

Another way to bring out the worry is by noting that encoding is entirely unconstrained. An object can encode, say, *being maroon* without thereby encoding *being red*. Absent this flexibility, the theory could not deliver impossible worlds. But this is further evidence that *encoding F* is not a viable reading of 'is *F*'. For if something is maroon, there's no escaping that it's red.

There are also worries about the notion of encoding itself. Byrd (1986, 247) argues that Zalta's 'dual predication view must face the task of giving a satisfactory account of the notion of encoding'. He questions whether a 'non-pictorial understanding can be articulated of the conditions under which '*o* encodes *F*' is true' (1986, 247). But that seems unfair to Zalta. *Encoding* is a primitive of his theory, so one can hardly ask for a definition. The choice of primitives is vindicated if the theory's successful applications reach far enough, and Zalta's approach certainly has wide scope (Zalta 1983, 1988).

A more pressing concern is that the facts about what properties an object encodes seem entirely ungrounded. When an object encodes, say, *being red*, there is no reason why it does so. It just does. We cannot say that it encodes *being red* because it encodes *being maroon*, or *being scarlet*, or some other determinate of red. It may encode no determinate of red at all. It is a primitive, unanalysable fact that that object encodes *being red*.

Similarly, when a world encodes *that A*, there is no reason why it does so. It's a metaphysically basic fact about a world that it encodes what it does. It's not even that a Zalta world encodes *that* $A \wedge B$ in virtue of encoding both *that A* and *that B*, since some (impossible) worlds encode *that* $A \wedge B$ but neither *that A* nor *that B*.

Zalta's approach to worlds seems to be a form of *primitivism* (which we'll discuss in §2.7). On such accounts, there's no story to

be told about why they represent (or encode) what they represent (or encode). They just do, and that's all we can say on the matter. We'll offer objections to that approach in §2.7. Zalta's approach adds a further worry, in that it's more theoretically complex that straightforward primitivism. It involves postulating the notion of *encoding*, as well as a claim about 'is' being ambiguous. But if this approach to worlds is at bottom a primitivist one, what would be lost, specifically in the theory of worlds, by simply stipulating that some worlds exist, which primitively represent that such-and-such?

2.7 Primitivism about Worlds

Primitivism about worlds says that there's no informative answer to the question, 'how do worlds represent?' Rather, worlds represent what they represent, and that's all we can say on the matter. There is no further theory to give of how or why a given world represents what it represents. Facts such as *that world w represents that A* are primitive facts: they are 'metaphysical bedrock'.

Primitivism about worldly representation is a natural match for primitivism about the metaphysical structure, or nature, of worlds. On this view, worlds have no analysable structure. They don't have parts, or constituents. We might think of them as dimensionless 'points' in modal space. This is the view Lewis (1986b, §3.4) calls 'magical ersatzism'. (See also Lewis 1986a and Nolan 2005 for discussion.)

One might go for primitivism about worlds because one thinks questions about their nature (and about how and why they represent what they do) are bad questions. But we don't think this line is very plausible, given willingness to analyse metaphysical notions of possibility and necessity. We can offer all kinds of explanations about how, for example, natural languages, paintings, photographs and the like represent what they do. In general, the question, 'how (or why) does X represent that *A*?' is meaningful. Why should the case of worlds be any different?

One might find motivation for primitivism about worlds from *quietism* about certain metaphysical questions. Quietists about a certain topic or question, such as moral truth or the ultimate nature of reality, don't look to provide a positive characterization of their subject matter. They deny that any such positive characterization is called for. That view naturally aligns with certain kinds of pragmatism (Macarthur 2008), although its supporters often find their inspiration in Wittgenstein (1953).

We don't think committed quietists should be primitivists about worlds. We're interested in the metaphysics of worlds because we're attempting to give a constructive analysis of metaphysical notions, including possibility and necessity, and perhaps of various notions of content, too. Neither pragmatist nor Wittgensteinian quietists will be attracted to this kind of constructive theorizing about those notions. Having gone in for constructive metaphysical theorizing, it's no good to then claim 'quietism!' whenever a tricky question arises.

A better motivation for primitivism, in our view, comes from the problems encountered in trying to answer the question, 'how does a world represent?' If none of the accounts surveyed above manage to give a satisfactory, informative answer, then primitivism becomes more attractive. On this approach, the question makes good sense, but has no informative answer.

Merricks (2015) offers an approach along these lines, but for propositions, not worlds. His view is that propositions are primitive, unanalysable entities, whose nature is to represent states of affairs, but for which there's no informative theory to be given of why they represent what they do. That's just what the primitivist about worlds says about worlds. Merricks's argumentative strategy is to consider each informative proposal (about propositions) in turn, arguing that all have irreparable problems. Primitivism then emerges that the last theory standing.

This comparison with Merricks's primitivism about propositions throws up an unexpected issue for the primitivist about worlds. Many of the issues we've encountered in our discussion of worlds have parallels in the literature on propositions. In particular, questions

about their constituents or structure, about how they represent what they do, and how these two facets interrelate, are central when discussing both worlds and propositions. Merricks's arguments in favour of primitivism about propositions, if they work at all, would seem to work just as well for primitivism about worlds.

So should one be a primitivist just about worlds, or just about propositions, or about both? Given primitivism about propositions (including facts about their modal and entailment properties), we have all we need to understand worlds in terms of propositions. But equally, given primitivism about worlds, we have all we need to understand propositions in terms of worlds. Why prefer one approach over the other? And yet being a primitivist about both worlds and propositions begins to look like a phobia against giving informative answers.

Our main issue with primitivism about worlds emerges from the following considerations. Take a complex representation, *that* $A \wedge B$. Primitivism, taken literally, denies any links between a world's representing *that* $A \wedge B$, representing *that* A, and representing *that* B. But that seems absurd. In order to theorize at all, we need to establish some link between conjunctions and their conjuncts. After all, conjunctions entail their conjuncts. That is a modal fact, and so needs to be accounted for in our theory of worlds.

A natural reply is that a world represents *that* $A \wedge B$ because it represents both *that* A and *that* B individually. But this forgets about impossible worlds. It is impossible for a conjunction, but not its conjuncts, to obtain. So we might infer that there's an impossible world representing *that* $A \wedge B$ but not *that* A. How so? The question is hard indeed, and it calls out for an answer.

To give a feel of what an answer might look like, we might claim that worlds somehow contain propositions, and so represent *that* A when they contain that very proposition. An impossible world representing *that* $A \wedge B$ but not *that* A would then be one which contains the conjunctive proposition but not its conjuncts. We mention this not because we think it best answers the question, but because it shows what an answer may look like. Our point is

that, faced with the question of complex representations, we may begin to doubt the very coherence of impossible worlds. We require a theory of how they can represent (e.g.) conjunctions but not the conjuncts. Replying 'they do so primitively' clearly will not assuage those worries.

Before moving on, we should note that, in a sense, the primitivist approach to worlds may be the implicit default in formal worlds-based semantics as practiced by many logicians, linguists, and computer scientists. They often endorse a practical instrumentalism, of the kind we hinted at at the end of §1.4, about what worlds are and how they represent. Worlds are treated as primitive points; a model is built by assigning atomic sentences randomly to worlds; and that's it. That's all fine, in a context where the aim is to investigate the logical properties of classes of such models. We shouldn't have to answer all the metaphysical questions before doing any semantics: that would be hugely impractical. However, we should not confuse this attitude with an *argument* for metaphysical primitivism. The semantic approach will work (just about) whatever the metaphysics says about worlds, and so neither requires nor implies metaphysical primitivism.

2.8 Fictionalism

We'll close this chapter with an anti-realist take on worlds, represented here by *fictionalism about worlds*. Fictionalism gives a negative answer to our initial question on REALISM. Recall (§2.1) that we're using 'realism' to mean the view that we can talk about, refer to, and quantify over worlds, even when we're being at our most metaphysically serious. We're using 'anti-realism' for the view that we may use world-talk, but without ontological commitment to non-actual worlds.

Fictionalist strategies in ontology and metaphysics have gained popularity in recent years. (Berto and Plebani (2015, chapter 6) survey different approaches.) According to such strategies, claims

that seem to commit us to entities that may be, for whatever reason, controversial – abstract objects, mathematical entities, propositions, properties (or properties of a certain kind) – can be, as the fictionalist motto goes, 'good without being true' (Field 1989). Numbers, for instance, are useful for counting objects. If there are *n* things of a certain kind *F*, it is useful to speak as if there is a number, *n*, that counts them. Such talk can help us as a representational aid, or to shorten and facilitate communication and inference, although what it appears to be about is really no part of the furniture of the world.

To see how this might be possible in general, let's switch the topic to moral properties as an example. Anti-realists in this area will typically say something like the following:

> There are moral truths, such as *it is wrong to cause unnecessary harm intentionally*. However, literally speaking, there are no moral properties: it is literally false that there is such a thing as the property of *being morally wrong*. But we can engage in a moral fiction, according to which there are such properties. While it is literally true that causing unnecessary harm is morally wrong, it is only true in the fiction that the property of moral wrongness is possessed by such acts.

On this way of thinking, the moral anti-realist pretends ('in the fiction') that reality is as the moral realist says it is. When the fiction delivers a moral verdict, say that such-and-such is wrong, she treats those actions as being wrong. But all the prima facie ontologically committing claims (such as 'there is a property of moral wrongness') are true only within the fiction. Since being true within the fiction is not factive (it does not export to being true *simpliciter*), the anti-realist doesn't feel any pressure to commit to such properties.

The anti-realist may then align 'everyday' moral talk with her 'in the fiction' talk, and reserve her 'out of the fiction' talk for the kind of debate we typically engage in when doing metaphysics. (Note that this fictionalist isn't an *error theorist* about morality, in Mackie's (1977) sense. For, unlike Mackie, she allows for moral truth.)

This kind of move is appealing when the realist's metaphysics seems to deliver good results, but you can't quite bring yourself to accept

her ontological claims. So, for example, if you agree with Mackie that moral properties would be 'a very strange sort, utterly different from anything else in the universe' (1977, 38), then it's tempting to be a fictionalist about moral properties.

Similarly, you can engage in *modal fictionalism* (Rosen 1990). The name is a bit misleading: modal fictionalists are not fictionalists about modality, but rather about *worlds*. According to fictionalists, talk of (and quantification over) worlds should be understood as literally false: it is only true within a 'worlds fiction', which we make-believe because it gives useful results in the analysis of modal notions. (For a survey, see Nolan 2016a.)

To understand how the fiction might work, take the following passage:

> There are countless other worlds … The worlds are something like remote planets; except that most of them are much bigger than mere planets, and they are not remote. Neither are they nearby. They are not at any spatial distance whatever from here. They are not far in the past or future, nor for that matter near; they are not at any temporal distance whatever from now. They are isolated: there are no spatiotemporal relations at all between things that belong to different worlds. Nor does anything that happens at one world cause anything to happen at another. Nor do they overlap; they have no parts in common. … The worlds are many and varied. There are enough of them to afford worlds where (roughly speaking) I finish on schedule, or I write on behalf of impossibilia, or I do not exist, or there are no people at all, or the physical constants do not permit life, or totally different laws govern the doings of alien particles with alien properties. There are so many other worlds, in fact, that absolutely every way that a world could possibly be is a way that some world is. (Lewis 1986b, 1)

You might be attracted to this 'philosophers' paradise' of genuine Lewisian worlds, while being unable to bring yourself to believe in the real existence of talking wombats and other non-actual entities. Then it's tempting to be a fictionalist about worlds, interpreting these Lewisian claims as true only within a worlds fiction.

Modal fictionalism, as presented in Rosen (1990), has been understood as fictionalism about possible worlds. If it works, it seems that its extension to impossible worlds should be straightforward. After all, if modal fictionalism is right about possible worlds, there really are no such things (other than the actual one). One may add that it's also true in the fiction that there are impossible worlds – but no worries, for in reality there are no such things. Similarly, it would have done no harm to *The Lord of the Rings* if Tolkien had added winged goblins to the population of Middle Earth, had the addition been useful to the overall plot.

How should we choose the right modal fiction? This is an issue raised by Sainsbury (2010, chapter 8). In the case of other forms of fictionalism, such as the one embedded in Field's nominalistic reconstruction of mathematical discourse (Field 1980, 1989), there is one standard story, namely standard mathematics. In this case, we have a conservativeness constraint with respect to such story: only a fiction which is conservative with respect to the results of standard mathematics will be acceptable. In this case, established mathematical results provide the relevant constraints.

In the modal case, however, there is no standard worlds-story to tell. Fictionalism about worlds is probably most attractive when it takes a genuine realist account as its fiction. Rosen (1990), for example, formulates his modal fictionalism by taking Lewis's *On the Plurality of Worlds* (1986b) as the relevant fiction. But what is the standard fiction for impossible worlds? Perhaps it is Yagisawa's 2010; perhaps Priest's 2016b. (The latter would be a somewhat strange choice for modal fictionalists, resulting in a fiction according to which some entities don't exist!)

The point we wish to stress is that fictionalism, in and of itself, doesn't propose any solution to the issues genuine realism faces with impossible worlds (§2.2). If those problems render genuine realism problematic, or even inconsistent, then we can't base the fiction of worlds on genuine realism (Jago 2016, §8). An adequate account requires the fictionalist first to solve the genuine realist's problems with impossible worlds. That is no easy task.

Fictionalists aim to get the advantages of realism about worlds, but without the ontological costs. Yet it is sometimes hard to tell just what the fictionalist's ontological beliefs are. She analyses modal talk based on what is true in the fiction. But within the fiction, worlds exist (in the most literal sense). So how is she to avoid saying that worlds exist, simpliciter? It might be that she uses her 'in the fiction' operator only for claims beginning 'it is possible that …' or 'it is necessary that …'. But 'it is necessary that there exist worlds' is true in the fiction and hence, she is forced to say, true simpliciter. But this entails that worlds exist! Fictionalists have to work hard to avoid this issue.

Another issue with fictionalism about worlds is that, if the worlds do not literally exist, then neither do constructions out of worlds. World-based theories often analyse propositions, meanings, subject matters, and other notions of semantic content, in terms of constructions from worlds. The fictionalist can do this too, in her fiction, but must say that, literally speaking, there are no such entities as propositions, meanings, or subject matters.

This may be a worrying conclusion to draw. If truth is a property which attaches to propositions, as many philosophers hold, but there are no propositions, then there are no truths. That can't be right. Similarly, some sentences have the same meaning: 'raccoons like to somersault' and 'procioni piace capriola' have the same meaning. But if they have the same meaning, then they have a meaning. Some thing is their meaning, shared between the two of them. This puts fictionalism about worlds in an awkward position. Either they deny that there are propositions, meanings, and so on, or else they accept that they exist and analyse them in some other way, without using worlds. But this undercuts much of the motivation for talking about worlds in the first place (§1.2).

Chapter Summary

The metaphysics of possible and impossible worlds revolves around a number of questions (§2.1). Should we treat worlds as *genuine* entities, which represent something as being *F* by having an *F* as a part? This is a hard position to maintain in the case of impossible worlds (§2.2). Should we treat (non-actual) worlds as non-existent beings (§2.3)? Or should we think of them as abstract entities and, if so, what kind of abstract entity (2.4)? Should we give the same answer to these questions for possible and impossible worlds (§2.5)? Yet a further option is to distinguish two senses of 'is' – *encoding* vs *exemplifying* some property – and claim that impossible worlds encode without exemplifying impossibilities (§2.6).

We argued that all of these approaches face difficulties. If one thinks these difficulties are insuperable, one can always adopt the approach that worlds are primitive entities (§2.7). We argued that this is a difficult line to maintain. Another fallback position is *fictionalism* about world, on which truths about worlds are always given 'in the fiction' (§2.8).

Of all the views considered, the most promising seems to be the one which takes impossible worlds (and perhaps the non-actual possible worlds too) to be ersatz entities. There are many competing ways to understand this suggestion (§2.4). We'll devote the next chapter to discussing which, if any, of these approaches is the most promising.

3

Ersatz Modal Realism

3.1 Classifying Ersatz Theories

We are using 'ersatz world' in contrast to 'genuine world'. In §2.2, we understood a genuine world as one which represents the existence of an *F* by having a real *F* as a part. So ersatz worlds are those that represent the existence of an *F* some other way.

That way of drawing the distinction between genuine and ersatz worlds cuts across the distinction between concrete and abstract worlds. Genuine possible worlds may line up with the concrete worlds, and ersatz possible worlds with the abstract ones. But when we want to talk about impossible worlds as well as possible worlds, these distinctions come apart (§2.2).

Our notion of an ersatz world is essentially a negative one: these are worlds which do not represent in the way genuine worlds do. So how do they represent? Different answers give us different ersatz theories. In §2.4, we introduced a number of ways (in general) to represent *that A*:

(State) By using a state of affairs of a certain kind.

(Property) By using a property reality would have, were things such that *A*;

(Combinatorial) By taking objects and properties which, if recombined in a given way, would make it the case that *A*;

(MAP) By using a map, picture, or image which depicts things being such that *A*;

(PROPOSITIONAL) By using the proposition or some other content-carrying entity *that A*;

(LINGUISTIC) By using bits of language, whose meaning is *that A*;

(PRIMITIVE) By taking the relevant representation to be a basic, unanalysable feature.

We've already discussed PRIMITIVE (in §2.7), which we treated as a theory separate from ersatzism. In the rest of this chapter, we'll discuss the remaining options. We'll explore how the various views work in the case of possible worlds, before investigating whether they can also accommodate impossible worlds. In several cases, we'll have to go beyond what the defenders of these views say, since they often don't explicitly discuss impossible worlds. We'll then indicate how we think the views in question may or should be extended to include impossible worlds.

Throughout our discussion, we'll talk of *worlds*, by which we'll normally mean worlds other than the reality around us. As explained in §2.4, ersatz theorists normally include an actualized ersatz world, which represents things being as they in fact are. This move allows them to analyse modal notions purely in terms of ersatz worlds. (Without an ersatz actual world, they'll need to say that *A* is possible iff there's an ersatz-or-genuine world such that *A*, which is a bit clunky. But nothing really hangs on this.)

3.2 Maximal States of Affairs

Plantinga (1970, 1974, 1976) develops a view on which possible worlds are maximal possible states of affairs. A state of affairs is maximal when it either includes or precludes every other state of affairs. Here, both *inclusion* and *preclusion* are understood modally: *s includes t* when, necessarily, if *s* obtains then so does *t*; and *s precludes*

t when, necessarily, if *s* obtains then *t* does not. Intuitively, a maximal state of affairs could not have any states of affairs added to it, without becoming impossible.

This approach, then, makes no attempt at a reductive definition of possibility or necessity. To understand 'maximal state of affairs', we first need to understand which states of affairs can obtain with which, and which they cannot obtain without. Modality is taken as a primitive concept.

A key idea is that a state of affairs can exist without *obtaining*. The merely possible worlds are those maximal states of affairs that exist, but do not obtain. Our world – the reality around us – determines the unique obtaining maximal state of affairs. Non-obtaining states of affairs are 'ontologically inert', in the sense that, if the state of affairs *that Fa* does not obtain, then it is not the case that *a* is *F*. We'll discuss just what this might mean at the end of this section.

As a consequence, the approach is *actualist*: all the states of affairs it posits actually exist. When we say 'non-actual world', we mean a state of affairs which exists, but does not obtain. Non-actual worlds represent an *F* by including a state of affairs *that a is F*, for some *a* or other. Such an *a* is represented as being *F* by that state of affairs, and hence by that world. But if that state of affairs does not obtain, *a* is not *F*. That's why we're classifying this view as an ersatz (non-genuine) account of worlds.

Can this approach accommodate impossible worlds? As far as we know, there is little discussion on this in the literature, one exception being Vander Laan 1997. To investigate the issue, we need first to amend the initial approach. Recall that maximality is defined in terms of either including or precluding every other state, with both notions defined modally. Now suppose we add a contradictory state of affairs, *s*, which represents that $A \wedge \neg A$. As *s* cannot possibly obtain, by definition it precludes all states of affairs (including itself), and includes no state of affairs (not even itself). It is maximal, but certainly should not be counted as a possible world.

We do better by thinking of states of affairs structurally, rather than (just) modally. Let us suppose we have some kind of grouping

operation on states of affairs, which makes a plurality into a unity. Think of this as a conjunction operation: if states $s_1, \ldots, s_n$ are grouped as an entity s, then s is the conjunctive state of affairs whose conjuncts are s_1 to s_n. We might, but need not, treat s as a sort of mereological sum or fusion, $s_1 \sqcup \cdots \sqcup s_n$.

This operation would give the notion of inclusion we need: the conjunctive state includes its conjuncts, but no more. Next, take a pair of states s_1, s_2 to be *incompatible* when it is not possible for them to obtain together. A maximal state is one which, for every incompatible pair of possible states s_1, s_2, includes either s_1 or s_2. (Alternatively, we could work with both negative and positive of states of affairs, as Barker and Jago (2012) do, and take a maximal state to be one which includes one state of each pair, *that A*, *that* $\neg A$.) We then identify the possible worlds with the maximal possible states of affairs, and the impossible worlds with all other states of affairs.

There are thus two ways in which a state of affairs may count as an impossible world. It may be a state that could not possibly obtain (an impossible state), or it may be a non-maximal state. (It may be both.) The former are (metaphysically) inconsistent worlds, the latter (metaphysically) incomplete worlds. We should flag that, on these definitions, an incomplete world may obtain, as part of a larger possible world. The state of affairs *that Charlie's tail is wagging* obtains, as part of the actual world, and hence is possible. But we nevertheless identify that state with an incomplete world, because it could not possibly be all there is to reality.

(We could alter our definitions so that only *non-obtaining* non-maximal states count as incomplete impossible worlds. Then, it would be a contingent matter whether some state s counts as a world. That seems strange to us. But since the matter is largely definitional, nothing much hangs on it.)

What does it mean to say a state of affairs exists, but does not obtain? We can't make much sense of the idea. The primary metaphysical role for states of affairs is in accounting for the ontology of predication. States of affairs provide an answer to the question, 'what is it for particular a to possess property F?' We might

understand the state of affairs *that a is F* as being (in some way) composed of *a* and *F* (Armstrong 1997, 2004). When property and particular are composed in that way, *a* is *F*. That's precisely what it is for *a* to possess *F*. But then, the state of affairs *that a is F* cannot exist without *a* being *F*. So, on Armstrong's approach, a state of affairs obtains just in case it exists.

That argument relies on Armstrong's specific theory of states of affairs. But in general, it is hard to see how some other approach would avoid this problem. If one determines the identity and existence of *that a is F*, but allows that *a* is not *F*, then one undermines the main argument for believing in states of affairs.

There is an independent argument against allowing states of affairs to exist without obtaining (Jago 2018b, §6.2). Those who believe in states of affairs usually take them to be *truthmakers* for the corresponding propositions. The truthmaker for the proposition ⟨*a* is *F*⟩ is the state of affairs *that a is F*. Usually, this is understood as the claim that the existence of the state of affairs *that a is F* is what makes ⟨*a* is *F*⟩ true. But we cannot say that if states of affairs may exist without obtaining. If *that a is F* exists but does not obtain, then *a* is not *F*, so ⟨*a* is *F*⟩ is not true and hence has no truthmaker.

Instead, the defender of non-obtaining states of affairs will say that a state of affairs makes a proposition true only when it obtains. But this move is problematic. Take a statement of the view in question,

(3.1) There exist states of affairs which do not obtain.

As an existential claim, this is made true (if it is true) by the very things it claims to exist. (Just about everyone agrees with this, even if they do not subscribe to truthmaker theory in general.) The only candidate truthmakers for (3.1) are states of affairs which do not obtain. But this conflicts with the principle above, on which non-obtaining states of affairs do no act as truthmakers. So, we do not think existent but non-obtaining states of affairs are coherent.

Overall, we don't see much hope for overcoming these issues. The notion of non-obtaining states of affairs is absolutely central to the

approach. So if it can't be put on a good metaphysical footing – and we don't see how it can – then the approach is doomed. However, there are other approaches in the vicinity, the propositional and linguistic approaches, which seem to capture the advantages of the states of affairs approach. We'll discuss these in §3.6 below.

3.3 Property Ersatzism

Possible worlds are ways that things could be or have been. Ways are properties. Thus, possible worlds are properties. So says *property ersatzism*. (This view is sometimes called *Stalnakerian realism* or *Stalnakerian ersatzism*, after Stalnaker (1976a); see also Forrest 1986 and Bigelow and Pargetter 1990.) A (non-actual) possible world is identified with the property that would be instantiated, were that world actualized. Such properties must entail a specific pattern of property-instantiations. That is, instantiating a world-property must entail all the specific matters of fact associated with the world in question. As Divers (2002, 177) notes, the approach hasn't been developed greatly, even in the case of possible worlds. So some of what we present in this section is our take on the most promising way to develop the approach.

We might think of a world-property as a big distributional property. Suppose we think of reality in the Humean way, as a pattern of matter, or point-sized particles, scattered over spacetime (Lewis 1986b). Then, a world-property would be a property which specifies a total distribution of matter (or point-sized particles) over spacetime.

This approach to world-properties is rather limited, however. It makes the assumption (usually called *Humean supervenience*: see Lewis 1994) that, given the facts about how matter is distributed locally, all the other facts follow. That's questionable, even assuming classical physics, and probably wrong given what we know about quantum physics. (Ladyman and Ross (2007) argue at length against the compatibility of Humean supervenience with quantum physics.)

We mention the idea merely to illustrate what a world-property might look like.

An alternative, and more flexible, approach is to take world-properties to be big conjunctive properties. We can make sense of the idea in terms of an operation on properties, forming a unified property from a plurality. (In §3.1, we mentioned a similar operation on states of affairs. As in that case, the operation on properties may, but need not, be identified with mereological sum or fusion. If we make that identification, then conjunctive properties are fusions of their conjuncts.)

On this approach, the conjoined properties are instantiated by whatever instantiates their conjunction: something possesses *being both F and G* when that very same thing is *F* and also *G*. World-properties are instantiated by reality as a whole, when they are instantiated at all. So, on this view, each conjunct of the conjunctive world-property must be a property possibly possessed by reality as a whole. These will be properties like *being such that Charlie is wagging her tail*. If instantiated, the corresponding state of affairs, *that reality is such that Charlie is wagging her tail*, straightforwardly entails the state of affairs *that Charlie is wagging her tail*.

The details of this approach will be much as they were for the maximal states of affairs account (§3.1). The difference here is that we identify non-actual worlds with uninstantiated properties, rather than non-obtaining states of affairs. On this approach, worlds are actual when they are instantiated (rather than when they obtain). A key advantage of this approach is that it is easier to make metaphysical sense of uninstantiated properties than it is of non-obtaining states of affairs.

When we turn to impossible worlds, the moves we made in §3.1 are attractive options here too. A property *F* is maximal when, for each incompatible pair F_1, F_2 of properties possibly possessed by reality, *F* includes either F_1 or F_2 as a conjunct. (Alternatively, we might work with negative properties, taking F_2 to be the negation of F_1.) We then identify possible worlds with maximal possible properties of reality. Impossible worlds are identified with all other properties of

reality: those that are non-maximal, plus those that reality could not possibly instantiate. (The approach is thus reductive about worlds, but non-reductive about which worlds are possible.)

Understood this way, property ersatzism can be viewed a form of linguistic ersatzism (§3.6). Specific properties of reality, *being such that A*, can be thought of as sentence-like or proposition-like entities, carrying the content *that A*. They can be conjoined, and perhaps negated too. So we will defer further discussion of this approach until §3.6, where we will discuss the issues faced by linguistic ersatzism.

3.4 Combinatorial Ersatzism

Combinatorial ersatzism comes in various shapes, but all of its variants agree that possible worlds should be understood in terms of recombinations of bits of actual reality. These can be actual individuals and actually instantiated properties, or obtaining simple states of affairs. Examples of the strategy are found in Quine 1969 and Cresswell 1972. We will focus on Armstrong's (1989, 1997) exemplary version, which he develops in terms of recombinations of actual particulars and universals (properties and relations). Key to this idea is Armstrong's notion of *sparse universals*: fundamental properties like *charge*, *spin*, and *mass*, as opposed to derivative ones like *being a penguin* and *being an* X Factor *contestant*. We can't freely recombine any property with any particular. Electrons can't possibly be red, and you can't possibly have spin ½. But, says Armstrong, we can freely recombine the sparse properties with the simple individuals.

Armstrong's aim is to give a reductive account of modality. He spells out which (actual) particulars and universals may be freely recombined, and spells out what it is to recombine them. All of this is done without bringing in modal notions. Possible worlds are then identified with recombinations of the selected particulars and universals, and modality is analysed in the usual way in terms of possible worlds.

On this picture, each possible world consists of a rearrangement of fundamental bits of reality. But possible worlds represent more than that: they also represent facts about penguins, *X-Factor* contestants, and the like. On Armstrong's view, all the non-fundamental facts represented by a world come for free, given facts about what's fundamental (according to that world). What a possible world *is*, according to Armstrong, is a rearrangement of fundamental stuff. But what it represents is whatever follows, metaphysically, from those facts, rearranged thus-and-so.

There are a number of issues with the approach, which we'll mention only briefly here. First, it's clear that not every fundamental bit of reality exists necessarily. This or that quark (or whatever) need not have existed. So merely rearranging actual, fundamental stuff won't give the right results. Instead, we have to think in terms of rearrangements which potentially leave some bits out.

By the same token, it seems that there could have been (fundamental) things which don't in fact exist. There could have been more quarks (or whatever) than in fact there are. No rearrangement of actual stuff produces those non-actual entities. So we have to think in terms of rearrangements which potentially include extra stuff (as well as potentially leaving some bits out). Those extra bits are *alien entities*: entities which are possible, but which don't actually exist. They're a real problem for this approach (very likely, for any actualist approach). After all, what are we supposed to be arranging? If it's actual stuff, plus some extra bits, then we seem to have quantified over non-actual entities, and the actualist endeavour is over. Moreover, if we are allowed to quantify over non-actual entities, why not instead go for genuine modal realism (§2.2)? This is a serious problem, but it's a problem any ersatz account must face. We'll set it aside for now, and return to it in §3.7.

The issue we want to investigate here is: just what *are* recombinations, metaphysically speaking? We'll argue that the recombination approach is destined to collapse into one of the other approaches we've discussed. Suppose rearrangements are (or consist of) states of affairs. Do the non-actual states of affairs obtain or not? If we say

they don't obtain, then we have the states of affairs account from §3.1. If they do obtain (perhaps on the grounds that, for states of affairs, *obtaining* amounts to *existing*), then we have a form of genuine realism. (Or we may have no theory at all. If possible worlds consist of genuine states of affairs, then modal space as a whole will contain contradictory states of affairs. It's hard to see how this approach can avoid outright triviality, wherein every sentence is treated as being true *simpliciter* (Jago 2018b).)

Rearrangements might be treated not as states of affairs, but rather as ersatz replacements for them. These might be propositions, or some other kind of property-and-particular-containing representational entity. That approach will be a form of propositional or linguistic ersatzism (to be discussed in §3.6). Armstrong's view is that the rearrangements do not in fact exist; but it is nevertheless convenient and acceptable to speak as if they do (Armstrong 1989, 49–51). As Sider (2005) argues, this is a form of fictionalism (§2.8). Possible-worlds talk should be understood as talk about a fictional ontology of recombinations. So, however we understand the ontology of recombinations, we don't have a genuinely distinct account of what possible worlds are.

It's also worth flagging a specific difficulty for Armstrong's fictional recombinations. If the fiction talks of all the recombinations at once, then it's at serious risk of inconsistency. In that fiction, there will exist pairs of logically incompatible states of affairs. Their existence renders that fiction logically inconsistent, and this in turn renders the entire account inconsistent (Jago 2016). The problem is avoided if there's a separate fiction for each recombination. But then, there's no way to analyse iterated modality, as when we say, it's possibly necessary that such-and-such.

Combinatorialism isn't best viewed as a theory of what possible worlds are. Rather, it's a theory of the extent of the space of possible worlds. How do things change when we include impossible worlds? We are not aware of any proposals in the literature, and so what follows are our own suggestions. It seems clear that, if all the rearrangements correspond to possible worlds, then we

can't make sense of impossible worlds in terms of those same rearrangements. One option is to make space in the theory for rearrangements which aren't licensed by Armstrong's restriction to sparse universals. We allow the derivative properties to 'float free' of their fundamental grounds. So, rather than rearranging facts about tables by manipulating fundamental facts about (say) the arrangement of matter in spacetime, we could think in terms of the property *being a table* itself. We may rearrange properties to arrive at a (representation of a) massless table, for example.

This approach will generate (representations of) impossibilities. We can recover the possible worlds as those that conform to acceptable recombinations of Armstrong's sparse universals. To do that, we'll need enough *bridge principles*, from fundamental to derivative facts. These 'vertical' principles tell us how the derivative facts are *grounded* in the fundamental facts. They should tell us, for example, that there's no fundamental recombination to support our massless table, and hence that that recombination gives us an impossible world. In general, when one of these grounding principles is broken, we have an impossible world.

This notion of *grounding* – a form of metaphysical dependence, of the kind we briefly discussed in §1.3 – is a much-debated concept in contemporary metaphysics (Fine 2001, 2012b,c, Schaffer 2009). If sense can be made of the notion using non-modal vocabulary, then the approach suggested here may offer a reductive and actualist-friendly analysis of modality.

3.5 Map Ersatzism

Map erstzism takes worlds to represent in much the same way as a realistic painting, a map, or architect's scale model does (Lewis 1986b, chapter 3). Consider how a map works. It has a limited vocabulary of symbols, arranged spatially, in a way that mimics the spatial arrangement of the bits of reality thereby represented. Maps typically use *above* and *below* (on the map) to represent *north of*

and *south of* (in reality), so that a tree-symbol above a lake-symbol represents a situation in which there are trees to the north of a lake. The distance between the symbols on the map, together with the map's scale, represents the distance in reality between trees and lake.

Map-based representations have some very useful representational properties. Suppose, in our map with a tree-symbol above the lake-symbol, there's also a church-symbol under the lake-symbol. Then the map represents the trees as being north of the church. And it represents the church as being south of both the trees and the lake. It does this for free. If we'd represented the tree-lake and lake-church relations descriptively, using sentences, then we'd have to do some work to infer that the trees are north of the church. And similarly, if we'd represented those facts linguistically using 'north', then we'd have to do some further work to infer that the church is south of the trees. Map-based representations build this inferential work automatically into the way they represent.

This feature makes map-based representations great candidates where what is represented is closed under (some amount of) logical consequence. Block (1983), for example, argues that cognitive data suggests that (at least some) mental representations are image-like, rather than linguistic. (See also Kosslyn and Pomerantz 1977, Paivio 1986, and Pinker 1980.)

But by the same token, this feature makes it hard for map-based representations to represent certain *im*possibilities. Just take the impossible situation where the trees are north of the lake, but the lake isn't south of the trees. How could a map represent that? Escher was ingenious in finding ways to depict many impossible situations pictorially, but the applicability of these or any other pictorial techniques is fairly limited. How would a map or picture depict an explicitly contradictory situation, say the one described by Graham Priest in *Sylvan's Box* (Priest 1997b), in which the narrator discovers a box that is both empty and not empty? (We discuss *Sylvan's Box* in detail in §11.3.) How would we even begin to depict the impossible situation in which Fermat's Last Theorem turns out false?

There are options here, to be sure. A map may contain a special *not like this* symbol, with maps featuring that symbol interpreted as the global negation of what they would otherwise represent. We may further introduce a conjunction operation on maps, with the resulting map representing the conjunction of what each map individually represents. Conjunction plus negation will allow us to represent explicit impossibilities, for sure. But we get there using linguistic, not pictorial, techniques. It seems likely that any form of map or pictorial ersatzism will need to resort to non-pictorial, linguistic ways of representing, if it aims to represent all the desired possibilities and impossibilities. But then why not go for linguistic representation across the board, and do away with any worries about how the pictorial and linguistic elements are supposed to interact?

3.6 Propositional and Linguistic Ersatzism

Propositional ersatzism identifies worlds with sets of propositions. Linguistic ersatzism is similar, except it builds worlds from sentences of some chosen language, the *worldmaking language*, rather than propositions. Linguistic forms of ersatzism go back to Carnap's (1947) theory of *state descriptions*, and both Adams (1974) and Jeffrey (1965) talk of *world-stories*, understood as sets of propositions. (They don't have to be sets. They could be some other kind of construction from propositions or sentences. All that matters is that we can take all the propositions or sentences making for a world together, treating them as a single entity. We'll work with sets.)

One nice feature of such worlds is the definition of truth, relative to a world. It is true that *A*, relative to world *w*, just in case the proposition *that A*, or the worldmaking sentence which means *that A*, is a member of world *w*. In this kind of definition, we have a phrase 'that *A*' in both the definiens and the definiendum. But it isn't circular. We're defining the truth of an English sentence, relative to a world, either in terms of a proposition or of a sentence of some other

language, the worldmaking language. In the latter case, this amounts to a translation from the worldmaking language into English.

Just what properties worlds have depends on our theory of propositions or our construction of the worldmaking language. Let's consider propositions first. One prominent theory of propositions identifies them with sets of possible worlds (Lewis 1986b, Stalnaker 1984). But that approach is ruled out, if we first identify worlds with sets of propositions. Indeed, that is one motivation for rejecting propositional ersatzism altogether. Modal semantics often works with sets of worlds, and it's very natural to identify these sets with propositions, as we saw in Chapter 1.

Whether propositions really are sets of worlds is moot (Jago 2015, 2018b, Merricks 2015). An alternative is the *Russellian* account, after Russell (1904/1980), which identifies (atomic) propositions with ordered sequences (or tuples) of particulars, properties, and relations. King (1995, 1996, 2011), Salmon (1986, 2005), and Soames (1987, 2008) all defend a variant of this view. The proposition *that Charlie is wagging her tail* is the sequence ⟪*wagging*, Charlie, Charlie's tail⟫. (We'll use the notation '⟪ ⟫' for sequences and the more familiar '⟨ ⟩' for propositions.)

Logically complex propositions are treated as sequences which include (the semantic values of) logical connectives. We might identify these values with truth-functions, familiar from truth-table semantics (although we don't need to make this identification). Let NEG, CONJ, and DISJ be the truth-functions corresponding to *negation*, *conjunction*, and *disjunction*, respectively. Then we can treat negated, conjunctive, and disjunctive propositions as follows:

$$\langle \neg A \rangle = \langle\!\langle \textsc{neg}, \langle A \rangle \rangle\!\rangle$$
$$\langle A \wedge B \rangle = \langle\!\langle \textsc{conj}, \langle A \rangle, \langle B \rangle \rangle\!\rangle$$
$$\langle A \vee B \rangle = \langle\!\langle \textsc{disj}, \langle A \rangle, \langle B \rangle \rangle\!\rangle$$

Since order matters to sequences, the structure of these propositions is (relatively) fine-grained, and allows for distinct but logically equivalent propositions. Even very closely related propositions, such

as $\langle A \wedge B \rangle$ and $\langle B \wedge A \rangle$, will be treated as distinct entities, on this approach. We'll say more about the Russellian approach below.

Yet another approach is to take propositions to be metaphysically primitive, unanalysable entities (Merricks 2015). We discussed the corresponding primitivist view about worlds in §2.7. Many of the same considerations raised there apply also to primitivism about propositions, and so we won't discuss the view any further.

Let's turn to linguistic ersatzism. Linguistic ersatzism is similar to propositional ersatzism, but with sentences in place of propositions. Worlds can be taken as sets of sentences. A theory along these lines must first specify the language to which those sentences belong. This is what is called the worldmaking language. The choice of language will be important. To see why, suppose we adopted English as our worldmaking language. Then, a world w represents that A iff it contains the English sentence 'A'. And so, on this analysis, it is possible that A iff there's a possible world which contains the English sentence 'A'. To determine whether this is the case, we will need to ask, which sets of English sentences are *compossible* (that is, possibly true all together)? But answering this question will depend, in part, on whether 'A' represents a possible situation. So we seem to have a circular analysis of possibility.

To avoid circularity, the worldmaking language and the theorist's language (English, in this case) must differ. One option is to use a *Lagadonian* language (Lewis 1986b, 145–6, following Carnap 1947), in which each particular is a name for itself and each property and relation is a predicate denoting itself. Infinitary logical connectives allow for sentences of infinite length (Divers 2002, 180).

The approach achieves little (goes one objection) if it requires that we first analyse which sets of sentences count as possible worlds. We do not have a reductive account of modality, of the kind genuine realists claim to have (§2.2). This is one of two objections Lewis (1986b) raises against linguistic ersatzism. (We'll discuss his second objection in §3.7.) The approach applies equally to propositional ersatzism, if the propositions in question are fine-grained as on the Russellian account.

In response, a defender of propositional or linguistic ersatzism may say that she never intended her account to be fully reductive. Its benefits lie elsewhere, she may claim. Sider (2002), for instance, argues that

> Unless modal consistency can be reduced in some way, linguistic-ersatz worlds cannot be used in a reductive analysis of modality, on pain of circularity. But the linguistic ersatzist can accept this limitation. The reduction of worlds to language still has a point, for it allows us to reduce all talk of worlds—which runs far beyond that which can be said utilizing merely the modal operators—to talk of possibility and necessity. As for these, they may one day be reduced in some way that does not involve worlds, or they may remain primitive. (Sider 2002, 282)
>
> A reduction of talk of possibilia which employs primitive possibility and necessity is nevertheless valuable since talk of possibilia runs beyond what can be said in the language of quantified modal logic. (Sider 2002, 306)

That's a concessive response. A more ambitious propositional or linguistic ersatzer might attempt to offer a fully reductive account of modality, by saying, in non-modal terms, which sets of propositions or sentences count as possible worlds. The difficulty here is that non-modal notions appear to be either too narrow, or too broad, to pick out metaphysical modality. To see why, consider those sets of propositions or sentences that are both logically consistent and closed under (classical) logical consequence. These are *maximal consistent* sets, and correspond to (classical) logically possible worlds. But not all of these represent (genuine, metaphysical, etc.) possibilities. There will be worlds where this apple is both red all over and green all over, or where there are married bachelors; these things aren't logically inconsistent. Lewis (1986b) argues, on these grounds, that this purely syntactic way of delimiting possible worlds gets the modal facts wrong, by delivering too broad a conception of what's possible.

The ambitious, reductive approach is difficult. Nevertheless, here's an attempt, which builds on the approach sketched at the end of §3.4. (We'll focus now on linguistic ersatzism, but a propositional

ersatzer could say similar things.) The metaphysically possible worlds are those logically possible worlds (the maximal consistent sets of sentences) which respect *grounding relationships*. Let's unpack this. *Grounding* is a metaphysical dependency relationship (§1.3) which holds between facts or states of affairs (and perhaps between other kinds of entity), where the grounded state of affairs obtains *in virtue of* the ground. The fact that it's now 20°C in Nottingham grounds the (more general) fact that it's now hot in Nottingham. If physicalism about the mind is true, then the fact that Anna's brain is in such-and-such state grounds the fact that she's feeling hot. And that fact in turn grounds the fact that the proposition ⟨Anna is feeling hot⟩ is true.

A set of sentences S respects grounding relationships as follows. Suppose S contains sentences $A_1, \ldots, A_n$, which represent states of affairs $s_1, \ldots, s_n$. Suppose also that $s_1, \ldots, s_n$ together ground a further state of affairs s. Then any sentence A which represents s must be a member of S, too. The idea is then that, when such sets of worldmaking sentences are consistent and closed under logical consequence, they (and only they) count as possible worlds.

This approach will work only if *grounding* is itself a non-modal notion. As we saw in §1.3, Fine (2001, 2012b) argues persuasively that grounding is irreducible to modal notions. Grounding has modal consequences – if state s_1 grounds s_2, then it is necessary that s_2 obtains if s_1 obtains – but it is not itself defined in modal terms. Indeed, it might be that grounding is a primitive, undefinable metaphysical notion. One of us attempts a reductive, non-modal analysis of grounding in Jago 2018a. Whether this approach will ultimately work out is beyond the scope of our discussion here. But if it does, then a fully reductive account of *possible world* to non-modal notions is on the cards for linguistic (and propositional) ersatzism.

3.7 Alien Entities

All versions of ersatz realism face a worry with *alien entities*: things that do not (actually) exist, but could have. Franz could have had an

older sister (but, in fact, doesn't). How is his possible older sister to be represented by an ersatz possible world? One way is by naming her. Let's call her 'Franzista'. But how do we give that name meaning? Not by associating it directly with an individual (as we do at a naming ceremony, for example), since, by hypothesis, there actually is no such person. We could instead give 'Franzista' its meaning by associating it with the description, 'Franz's older sister'. But this gives us strange results. Franz might not have existed. So we should expect the following situation to be possible: Franzista exists, but Franz does not. So, in that possible situation, Franzista is not Franz's sister. But this situation is contradictory if 'Franzista' means *Franz's sister*.

A better approach is to forget about naming Franz's merely possible older sister, and instead *describe* our target possible world. We can characterize that world as follows: *there are distinct individuals* $x_1, x_2, \ldots$, *each distinct from all actual individuals, such that* x_1 *is such-and-such,* x_2 *is such-and-such, …,* x_i *is Franz's older sister, ….* Here, we describe a possible situation in which there exists some older sister of Franz's. We can allow that that person may have had no siblings herself in some other possible situation.

A problem remains. How are we to capture this further possible situation, in which Franz's (merely possible) older sister has no siblings? Call the possible world just described, containing Franz and a Franz-sister, w_1. Let w_2 be a Franz-less world, in which Franz's sister from w_1 has no siblings. What guarantees that we are talking about the same person, Franz's possible sister, in both worlds? For all we've said, nothing does. So we have no way to say, thinking about Franz's sister in w_1, that *she* could have had no siblings.

One response to the problem is to use *counterpart theory* (Lewis 1968, 1971). This is a feature of Lewisian genuine modal realism, though we did not get into it in §2.2. Counterparts capture *de re* ways an entity could have been. If x and y are counterparts, then x could have been like y, and vice versa. In our ersatz setting, we can describe some x_i in w_1, and describe some x_j in w_2, and then say that x_1 in w_1 and x_j in w_2 are counterparts. One problem here is that the counterpart relation is, well, a relation, and relations relate entities.

Unless one is a Meinongian, if it's true that x_i and x_j are related, then x_i and x_j exist. (Note that the theory doesn't say that x_i and x_j are counterparts *according to world* w_1 or any other world. It's a claim of the theory from the viewpoint of all modal space.) But, by hypothesis, neither x_i nor x_j exist. So this can't be the right approach.

The problem of aliens gets even tougher to solve when we turn our attention to merely possible properties (Lewis 1986b, chapter 3). Presumably, there are properties that could have been instantiated but are not. Perhaps there are fundamental uninstantiated properties. They pose a serious problem for any ersatz approach. At least we have the means to describe Franz's merely possible older sister. We do so using the predicates 'older than' and 'sister of', which get their meaning through being associated with the corresponding relations. But in the case of alien fundamental properties, we don't even have the vocabulary to describe them.

Sider (2002) shows how to solve these problems. On his approach, rather than describing each possible world one by one, we describe all of them in one go. (For simplicity, we'll consider the case of alien individuals only.) We say: *there are distinct individuals* $x_1, x_2, \ldots$, *each distinct from all actual individuals, and worlds* $w_1, w_2, \ldots$, *such that, in* w_1, ..., *and in* w_2, ..., *and* .... Here, we quantify over merely possible individuals all in one go, and then describe how each is in each of the worlds which represent that individual. This is the *Ersatz pluriverse sentence*. It is false, of course, since (by actualist lights) there are no merely possible individuals. Its function is to represent a space of possibilities, just as (standard) linguistic ersatzism does. The ersatz pluriverse sentence makes clear when distinct worlds w_i and w_j are talking about the same possible particular. That is the great advantage of Sider's approach. Jago (2013c) raises a worry for Sider's approach, which we won't discuss here. This said, we conclude by saying that we take a form of linguistic ersatzism endowed with grounding relations to be the most promising ersatz account. One advantage of linguistic ersatzism is that it helps with an insidious objection to impossible worlds: the *compositionality objection*, which we discuss in §8.5.

Chapter Summary

We can understand ersatz possible worlds as maximal states of affairs; maximal properties; recombinations of actual bits of reality; as maps; or as entities built from propositions or sentences (§3.1). Our question was: can these approaches be extended to include impossible worlds? The states of affairs approach can, with some modification, accommodate impossible worlds; yet we found the concept of a non-obtaining state of affairs hard to sustain (§3.2). The property approach too can, with some modification, be extended to impossible worlds. We then argued (§3.3) that the extended approach is best viewed as a form of linguistic ersatzism.

The combinatorial faces the question: what *are* recombinations, metaphysically speaking? We argued that, however the question is answered, this approach collapses into one of the others (§3.4). Map ersatzism does not seem general enough to accommodate all the impossibilities our theories require of impossible worlds (§3.5). The most promising approach seems to be propositional or linguistic ersatzism, of which, we prefer the linguistic variant (§3.6). Finally, we discussed an issue all ersatz accounts face: the problem of *alien entities* (§3.7).

Part II

Logical Applications

4

Modal Logics

4.1 Normal Modal Logics

We begin with a rehearsal of standard or normal modal logic. To keep things simple, we limit ourselves to a propositional language $\mathscr{L}$, including a set of atoms AT: $p, q, r, p_1, p_2, \ldots$. We have negation $\neg$, conjunction $\wedge$, disjunction $\vee$, the material conditional $\supset$, and the box $\Box$ and diamond $\Diamond$ of necessity and possibility. We use $A, B, C, \ldots$ as metavariables for formulas of $\mathscr{L}$. The well-formed formulas are the atoms in AT and, if A and B are well-formed formulas, then so are:

$$\neg A \mid (A \wedge B) \mid (A \vee B) \mid (A \supset B) \mid \Box B \mid \Diamond B$$

Outermost brackets are normally omitted.

A *normal possible worlds frame* or *Kripke frame* $\mathscr{F}$ for $\mathscr{L}$ is a pair $\langle\langle W, R\rangle\rangle$, where W is a set of possible worlds and $R \subseteq W \times W$ is a binary accessibility relation between them. A frame becomes a *model* $\mathscr{M} = \langle\langle W, R, v\rangle\rangle$, when endowed with a valuation function v. This assigns to each atom either the value 1 (true) or the value 0 (false) at a world. So we write '$v_w(p) = 1$' to mean that p is true at w, and '$v_w(p) = 0$' to mean that it is false there.

The valuation function v is extended to the whole language via the following recursive clauses:

(S$\neg$) $v_w(\neg A) = 1$ if $v_w(A) = 0$, and 0 otherwise.

(S$\wedge$) $v_w(A \wedge B) = 1$ if $v_w(A) = v_w(B) = 1$, and 0 otherwise.

(S∨) $v_w(A \vee B) = 1$ if $v_w(A) = 1$ or $v_w(B) = 1$, and 0 otherwise.

(S⊃) $v_w(A \supset B) = 1$ if $v_w(A) = 0$ or $v_w(B) = 1$, and 0 otherwise.

(S□) $v_w(\Box A) = 1$ if for all $w_1 \in W$ such that Rww_1, $v_{w_1}(A) = 1$, and 0 otherwise.

(S◇) $v_w(\Diamond A) = 1$ if for some $w_1 \in W$ such that Rww_1, $v_{w_1}(A) = 1$, and 0 otherwise.

Logical consequence or entailment '⊨', is defined as truth preservation at all worlds of all models (for any set of formulas Γ):

> $\Gamma \models A$ iff for all models $\mathscr{M} = \langle\langle W, R, v\rangle\rangle$ and all $w \in W$: if $v_w(A) = 1$ for all $B \in \Gamma$, then $v_w(A) = 1$.

For single-premise entailment, we will write $A \models B$ instead of $\{A\} \models B$. Logical equivalence, $A \dashv\models B$, is two-way entailment between A and B. Logical validity or logical truth, $\models A$, defined as truth at all worlds of all models, is a special case of entailment by the empty set, $\varnothing \models A$.

The semantics makes $\Box A$ equivalent to $\neg\Diamond\neg A$ and $\Diamond A$ equivalent to $\neg\Box\neg A$, as desired. It also validates the *Distribution principle* or *K-principle*:

(K) $\Box(A \supset B) \supset (\Box A \supset \Box B)$

The logic induced by the semantics (the set of valid sentences) is called **K**, after Kripke. This is the weakest *normal modal logic*. In the context of modal logic, 'normal' means that the logic includes all the classical tautologies plus (K), and is closed under *modus ponens* and the *Necessitation* rule:

(N) If $\vdash A$, then $\vdash \Box A$

(Be careful in how you read this rule. It doesn't say that A implies $\Box A$: that would trivialize modality by committing us to treating all truths as necessary truths. Rather, it says that, if A is a theorem of the logic/a logical truth, then so is $\Box A$.)

K is the base normal modal logic, in that its semantics puts no conditions on the accessibility relation R. If we impose some conditions on R, we obtain stronger normal modal logics. The normal modal logics obtained in this way contain all the **K**-theorems, plus some extra ones too. (In semantic terms, we get more entailments by putting further conditions on the accessibility relation R.) Table 4.1 shows the most well-known cases:

Axiom name	Axiom scheme	Frame condition
D	$\Box A \supset \Diamond A$	R is serial: $\forall x \exists y Rxy$
T	$\Box A \supset A$	R is reflexive: $\forall x Rxx$
B	$A \supset \Box\Diamond A$	R is symmetrical $\forall x \forall y(Rxy \rightarrow Ryx)$
4	$\Box A \supset \Box\Box A$	R is transitive $\forall x \forall y \forall z(Rxy \wedge Ryz \rightarrow Rxz)$
5	$\Diamond A \supset \Box\Diamond A$	R is euclidean $\forall x \forall y \forall z(Rxy \wedge Rxz \rightarrow Ryz)$

Table 4.1: AXIOM-FRAME CORRESPONDENCE

The logic **KTB** adds the T and B axioms, for example, and so corresponds to to the reflexive and symmetrical frame. Corresponding to the serial and euclidean frame is the logic **KD5**. And so on.

Whether we accept these additional axioms depends on how we understand the involved modalities. The D axiom says that what is necessary is possible, and this seems plausible on most readings of the notions of possibility and necessity. For D to hold, we need every world in W to access some world (that's Seriality). For if a world w accesses no world, given (S$\Box$), all formulas of the form $\Box A$ are true at w. And given (S$\Diamond$), all formulas of the form $\Diamond A$ are false at w, for there is no accessible world where A is true. Thus, D fails. (The label 'D' comes from 'deontic', inspired by the reading of necessity as 'it ought to be the case that' and of possibility as 'it is permissible that'.)

If the relevant necessity is factive, then the T axiom must hold: if A is necessary (at a given world w), then it should be true (at that

world). This won't hold in general without Reflexivity (every world is possible relative to itself). For without Reflexivity, it could be that A holds at all worlds accessible from w but not at w itself.

If the relevant necessity is unrestricted, then the 4 and 5 axioms look very plausible. If A is unrestrictedly necessary (or possible) in the relevant sense, *this* fact should not be contingent on anything and so it should be necessary as well. This may not be so for factive but restricted or relative necessities. Against 4, for instance, it might be physically necessary (determined by the laws of physics) that bodies do not accelerate through the speed of light; but its necessity may not be determined by the laws of physics. The 4 and 5 axioms are characteristic of two important normal modal logics, **S4** (= **KT4**) and **S5** (= **KT5**), due to C.I. Lewis (Lewis and Langford 1932), a founding father of modern modal logic.

If we read '$\Box$' as an epistemic operator expressing knowledge, it is doubtful that either 4 or 5 holds. In epistemic logic, 4 is called the Axiom of (Positive) Introspection, or *KK-principle*. It says that, if one knows that A, then one knows that one knows that A. One has perfect introspective access to what one knows. It seems, however, that this has counterexamples. Think of yourself panicking the night before the exam, but doing fine with your essay the day after. You may truthfully say: 'Yesterday I didn't know I had learned so much by studying, but today it turned out that I did'. You knew the answers, but didn't know that you knew them all.

The 5 axiom is even more suspect in an epistemic setting. It is equivalent to $\neg\Box A \supset \Box\neg\Box A$, which, in an epistemic setting, says: if one doesn't know that A, then one knows that one doesn't know that A. That doesn't seem at all plausible. For one thing, we often think we know things we don't in fact know. In those cases, we don't believe, and so don't know, that we don't know them.

As we shall see in Chapter 5, it is indeed doubtful that *any* normal modal logic can provide an adequate formal treatment of epistemic notions like knowledge and belief. This is due to the logical omniscience phenomena, which were introduced in Chapter 1, and to which we shall return in §5.1.

4.2 Non-Normal Modal Logics

This section expands on Berto and Jago 2018. Normal Kripke frames are celebrated for having provided suitable interpretations of different systems of modal logic, including **S4** and **S5**. Before Kripke's work, we merely had lists of axioms or, at most, algebraic semantics many found rather uninformative. Kripke also introduced *non-normal worlds* (Kripke 1965), in order to provide world-based semantics for modal logics weaker than the basic normal modal system **K**. These are *non-normal* modal logics, including C.I. Lewis's systems **S2** and **S3**.

Non-normal modal systems do not include the Necessitation rule:

(N) If $\vdash A$, then $\vdash \Box A$

As we said in §4.1, (N) holds in the weakest normal modal logic **K** and all its normal extensions. A semantic counterpart of (N) would tell us that if A is a logical truth, then so is $\Box A$. This principle cannot be avoided when $\Box$ is understood in line with (S$\Box$). For if A is a logical truth, then it is by definition true at all worlds of all models. So given any world w, A is true at all worlds accessible from w, so also $\Box A$ is true at w. Since this applies to any world of any model, $\Box A$ will thereby be a logical truth, too.

Non-normal worlds enter the stage in order to make (N) fail. Take the same language $\mathscr{L}$ of §4.1 and give it the following semantics. A *non-normal worlds frame* $\mathscr{F}$ for $\mathscr{L}$ is a triple $\langle\langle W, N, R\rangle\rangle$, with W the set of worlds and $N \subseteq W$ the subset of normal worlds, so that the items in $W - N$ are the non-normal worlds. R is as before. A frame becomes a non-normal model $\mathscr{M} = \langle\langle W, N, R, v\rangle\rangle$ when endowed with a valuation function v assigning truth values to formulas at worlds.

The truth conditions for the extensional logical vocabulary are defined as in §4.1. But we now take the clauses (S$\Box$) and (S$\Diamond$) to apply to normal worlds only. If $w \in W - N$, the clauses are:

(NS$\Box$) $v_w(\Box A) = 0$

(NS$\Diamond$) $v_w(\Diamond A) = 1$

At non-normal worlds, formulas of the form $\$A$, with $\$$ a modal operator, are not evaluated depending on the truth value of A at other (accessible) worlds, but get assigned their truth value directly. Specifically, all $\Box$-formulas are false and all $\Diamond$-formulas are true. In a sense, non-normal worlds of this kind are worlds where nothing is necessary and anything is possible. These worlds are deviant only in this respect: their behavior, as far as the extensional connectives are concerned, is quite regular. Notice also that, as is easy to check, (NS$\Box$) and (NS$\Diamond$) still deliver the equivalence of $\Box A$ with $\neg\Diamond\neg A$ and of $\Diamond A$ with $\neg\Box\neg A$.

Logical consequence or entailment is defined as truth preservation at all *normal* worlds in all models:

> $\Gamma \vDash B$ iff for all models $\mathscr{M} = \langle\!\langle W, N, R, v \rangle\!\rangle$ and all $w \in N$: if $v_w(A) = 1$ for all $A \in \Gamma$, then $v_w(B) = 1$

Logical validity is truth at all normal worlds in all models.

Restricting logical consequence and validity to normal worlds in this way is a common, though not universal, move in semantics that include non-normal or impossible worlds. The insight behind this comes from the second characterization of impossible worlds as 'logic violators' from §1.4: worlds where logic is different, or where the laws of logic fail. If this is the interpretation of the items in $W - N$, then we should not refer to, or quantify over, such worlds when we characterize logical consequence and validity. For these, we want to look only at possible or normal worlds: worlds where logic is *not* different.

This setting gives us a basic non-normal modal logic, which Priest (2008) calls **N**. If one adds the condition that R be reflexive, one gets C.I. Lewis's modal system **S2**. If one takes R to be reflexive and transitive, one gets **S3** (Kripke 1965).

This kind of semantics makes (N) fail. Take any classical propositional tautology, say, of the form $A \vee \neg A$. This holds at all worlds of all interpretations. Therefore, $\Box(A \vee \neg A)$ holds at all normal worlds of all interpretations, so $\vDash \Box(A \vee \neg A)$. But by (NS$\Box$), $\Box(A \vee \neg A)$ does

not hold in any non-normal world. So $\Box\Box(A \vee \neg A)$ is false at normal worlds that have access, via R, to any non-normal world. As there must be some such world in some model, we have $\not\models \Box\Box(A \vee \neg A)$.

Another welcome feature of this semantics is that it does not make the 'irrelevant' conditional $A \dashv\!\!\supset (B \dashv\!\!\supset B)$ valid. This is one of the 'paradoxes' of the strict conditional (§1.3). It fails in non-normal models once one reads $A \dashv\!\!\supset B$ as the necessitation of the horseshoe, $\Box(A \supset B)$. Then the paradox is $\Box(A \supset \Box(B \supset B))$. This is valid in **K**, but not in **N** (and extensions): take a world $w \in N$ accessing a non-normal w_1 where A holds but (since the world is non-normal) $\Box(B \supset B)$ fails.

Non-normal worlds semantics of this kind does not provide a systematic framework for dealing with all cases of irrelevance (nor was it intended to do so). Nevertheless, this way of handling one paradox of the strict conditional hints at a general strategy: make irrelevant conditionals fail by taking into account non-normal or impossible worlds, understood as worlds where logical truths such as $B \dashv\!\!\supset B$ may fail. We'll see more of this approach in Chapter 6.

The semantics for non-normal modal logics such as **S2** and **S3** is based on a valuation function which assigns the same truth value to all $\Box$-formulas (false) and all $\Diamond$-formulas (true) at non-normal worlds. We also mention the modal system **S0.5**, due to Lemmon (1957). This is a non-normal system whose semantics, initially provided by Cresswell (1966), includes non-normal worlds at which formulas that begin with a modal operator are assigned arbitrary truth values. The valuation function v treats modal formulas as atomic. (Interpretations for **S2** or **S3** are special cases of interpretations for **S0.5**: those cases in which the valuation function uniformly treats $\Box$-formulas as false and $\Diamond$-formulas as true at non-normal worlds.) This setting makes the inter-definability of $\Box$ and $\Diamond$ via negation fail.

4.3 Non-Uniform Truth Conditions

A key idea in impossible worlds semantics of various kinds is that certain complex formulas are assigned arbitrary values at non-normal worlds. They are, in effect, treated as atomic sentences. At worlds that behave this way, the syntax of a formula can be partly or wholly disregarded. As we shall see in Chapter 5 on epistemic logics, this insight can be fruitful, and is at work in semantic frameworks including non-normal or impossible worlds.

This approach requires the truth conditions of some operators not to be spelled out in a uniform way across worlds. In particular, at non-normal worlds complex formulas can have truth values assigned in a non-recursive way. But this approach raises worries. Fine (2019) claims that it is a 'theoretical virtue in itself' for a semantics to be uniform:

> we would like the compositional clauses for the logical connectives to be 'uniform' or non-disjunctive. … without uniformity, it is not even clear that we will have clauses for the logical connectives themselves as opposed to some gerry-mandered product of the theoretician's mind. Fine (2019, 1)

A similar worry is pressed in Williamson (2017).

It is not clear, however, why disjunctiveness would be a problem. The fact that a concept has a disjunctive characterization per se does not make the concept itself gerrymandered or gruesome. The notion *Australian citizen*, for instance, is obviously perfectly fine even though it works just as follows: x is Australian citizen iff either x was born in Australia or x has been naturalized. (The example is due to Priest (2005, 237).)

A more serious worry is lack of compositionality. If truth-at-a-world-conditions constitute meaning, and we want meanings to be compositional for them to be graspable by finite minds, then the truth-at-a-world-conditions of whole formulas should be given in terms of those of their subformulas. If there are infinitely many non-normal worlds in our setting, however, we may have no way finitely to do this.

We think this is a serious philosophical worry. We'll offer a philosophical response to it in §8.5, by showing how to get a compositional account of content involving impossible (non-normal) worlds. Here, our focus is on the logical applications of impossible worlds. We want our impossible worlds models to give us a notion of logical consequence that's useful for certain applications, for example. But that isn't to demand that they are all compositional. A model can be useful, even if it isn't capable of underpinning a theory of meaning for a language like English. So for pragmatic logical purposes, we're happy to dismiss the worry.

4.4 Non-Adjunctive and Non-Prime Worlds

This section draws on Berto and Jago 2018. Rescher and Brandom (1980) introduce impossible worlds of a different kind to the ones we just saw. (Their book's subtitle is *A Study in Non-Standard Possible-Worlds Semantics and Ontology*, but, as we shall now see, the worlds making their semantics non-standard deserve the label of impossible worlds.)

In their formal semantics, standard possible worlds are taken as maximal consistent aggregates of states of affairs. Their non-standard worlds are obtained combinatorially, by means of two recursive operations having standard worlds as their base. These are *schematization* ('∩') and *superposition* ('∪'). Given w_1 and w_2, a schematic world $w_1 \cap w_2$ is one at which all and only the states of affairs obtain which obtain both at w_1 and at w_2. A superposed (or inconsistent) world $w_1 \cup w_2$ is one at which all and only the states of affairs obtain which obtain either at w_1 or at w_2. With respect to the definitions listed in §1.4, Rescher and Brandom's inconsistent-superposed worlds are impossible worlds of the fourth kind, that is, contradiction-realizers: they can make both A and $\neg A$ true. (Just superpose a possible world, w_1, at which you are 1.70m tall, and another one, w_2, at which you are not.)

Unlike Kripke's non-normal worlds, which require different truth conditions only for the modal operators, Rescher and Brandom's non-standard worlds behave peculiarly with respect to extensional operators. The standard clause for conjunction (S$\wedge$) from §4.1 has to go for superposed worlds: A and B can each be true without $A \wedge B$'s being true. These worlds still have a certain amount of logical structure. They behave in quite a standard fashion with respect to essentially single-premise inferences (Priest et al. 1989, 161). But they are anarchic with respect to essentially multiple-premise inferences.

Dually, schematic worlds can be *non-prime*: it can happen that $A \vee B$ is true at a world without either A or B being true at that world. One application that can motivate non-prime worlds is the handling of under-determined information: one may have the information that Strasbourg is either in France or in Germany, without having information as to which is the case.

Now 'dualizing' back, one may use inconsistent-superposed worlds to model inconsistent databases. These may consist in sets of data or information, supplied by different sources which are inconsistent with each other, such as incompatible evidence presented by different witnesses in a trial. The non-adjunctive features of superposed worlds are useful here. Intuitively, one is allowed to draw the logical consequences of the data or information fed in by a single source, but one does not conjoin data from distinct sources which may be inconsistent with each other. The database is 'compartimentalized': occasional inconsistencies are placed in separate sectors, and not conjunctively asserted. This is an example of a *non-adjunctive* system. Hyde (1997), Lewis (1982), and Varzi (1997) each discuss different uses of this kind of approach. We will come back to uses of impossible worlds in information theory in more detail in Chapter 9.

Chapter Summary

After recapping standard normal modal logics and their frame correspondeces (§4.1), we introduced non-normal modal logics,

which invalidate the Necessitation rule (N). We showed how to model these logics using non-normal or impossible worlds, thought of as 'logic violators' (§4.2). This approach gives comes with non-uniform truth conditions: some operators are understood in one way at normal worlds, in another way at non-normal worlds (§4.3). This may or may not be a problem; we'll come back to it in §8.5. We then discussed the specific case of non-adjunctive and non-prime worlds (§4.4), where conjunction and disjunction can behave in unusual ways.

5

Epistemic Logics

5.1 Standard Epistemic Logic and Logical Omniscience

In §1.2, we introduced the idea of understanding knowledge and belief as restricted quantifiers over possible worlds, where the accessible worlds are those that represent epistemic possibilities for a cognitive agent. This can be modelled by taking the language $\mathscr{L}$ of §4.1 with its normal Kripke semantics, and interpreting '$\Box$' as an operator representing knowledge or belief. We'll rewrite '$\Box$' as 'K', for 'one knows that'. (We'll also talk about belief, and sometimes use 'B' in place of 'K'. But most of what we say about modelling knowledge goes for belief, and vice versa.) Its semantic clause is:

(*SK*) $v_w(KA) = 1$ if for all $w_1 \in W$ such that Rww_1, $v_{w_1}(A) = 1$, and 0 otherwise.

One can then read the dual '$\Diamond$' as 'it is compatible with what one knows/believes that'.

The relation R in a Kripke model $\langle\!\langle W, R, v\rangle\!\rangle$ should now be read in an epistemic way, as *epistemic accessibility*. Epistemic operators are often indexed to a particular agent, with 'K_iA' read as 'agent i knows that A'. Indexing allows a multi-modal logic representing the cognitive states of a plurality of agents. For agents $1, \ldots, n$, each gets its own knowledge operator $K_1, \ldots, K_n$, with a corresponding epistemic accessibility relation, $R_1, \ldots, R_n$.

It's sometimes useful to rephrase accessibility in terms of a function f which, given a world w as input, returns the set of worlds fw that are epistemically accessible from w. This f is called the *epistemic projection function*, and is defined by setting $fw = \{w_1 \mid Rww_1\}$. Then KA's truth at w requires A's truth at all worlds in fw. When we have multiple epistemic accessibility relations $R_1, \ldots, R_n$, we have multiple epistemic projection functions, $f_1, \ldots, f_n$, one for each agent. Since the issues we'll go on to discuss arise in both the single-agent and multi-agent settings, we'll focus for simplicity on single-agent models, with a single modality 'K'.

Our concern here will be with the various issues arising under the heading of *logical omniscience* (§1.3). The issues are the result of our semantics satisfying various *closure conditions*. These take the form: if an agent knows —, she must also know —. Fagin et al. (1995, 335–6) and Van Ditmarsch et al. (2008, 23) discuss the following (with '$A \dashv\models B$' meaning that $A \models B$ and $B \models A$):

(C1) If KA and $A \models B$, then KB

(C2) If $\models A$, then KA

(C3) $\models \neg(KA \wedge K\neg A)$

(C4) If KA, and $A \dashv\models B$, then KB

(C5) If KA and $\models A \supset B$, then KB

(C6) If KA and $K(A \supset B)$, then KB

(C7) If $K(A \wedge B)$, then KA and KB

(C8) If KA, then $K(A \vee B)$

(C1) is *Closure under entailment* or *Full omniscience*. (C2), *Knowledge of all valid formulas*, is a special case of (C1). We met both principles in §1.3. (C3), *Consistency*, says that one cannot know contradictory things. (C4), *Closure under logical equivalence*, follows from (C1) as logical equivalence is defined as two-way entailment.

(C5), *Closure under valid implication*, is equivalent to (C1) in systems in which $\vDash A \supset B$ if and only if $A \vDash B$. (C6) is *Closure under known implication*. (C7) and (C8) are often called *Closure under conjunction* and *Closure under disjunction*, respectively. There are corresponding closure principles for belief.

In §1.3, we discussed how (C1) and (C2) are implausible for the case of real, finite, and fallible cognitive agents. The corresponding principles for belief are just as implausible, as is (C3): we are all inconsistent believers. The other principles look no more promising. (C4) and (C5) obviously are in no better position than (C1). (C6) is a hotly contested principle in epistemology, because of its relation to external-world scepticism. You know that, if you have hands, then you're not a brain in a vat. But you don't know that you're not a brain in a vat. (How could you?) Given (C6), it follows that you don't know that you have hands (and similarly for other simple bits of external world knowledge). (Dretske (2005), Holliday (2015), Nozick (1981), and Roush (2010) discuss the issue.) (C7) is perhaps the most plausible principle in the list. As for (C8), it seems implausible that, just because one knows that A, one automatically knows the disjunction of it and any arbitrary B. One may lack the very concepts involved in B.

All of these principles except (C3) hold in the weakest normal modal logic **K** from §4.1, with no conditions on the accessibility relation R. ((C2) is related to (N) and (C6) to the K-principle, for example. (C3) requires Seriality to be valid.) This fact tells us that tampering with the accessibility relation is not going to help us avoid all of these principles. So if we want to understand knowledge and belief in terms of modal logic, we should not work with a normal modal logic.

Non-omniscience is often taken as evidence that knowledge and belief are hyperintensional. There are a number of ways to draw distinctions more fine-grained than those available in standard possible worlds semantics, however. Not all of those options resort to impossible worlds. We'll briefly discuss some of them in §5.2.

5.2 Dealing with Omniscience without Impossible Worlds

One radical way to avoid Logical Omniscience is to adopt a *syntactic approach* to modelling knowledge and belief. What an agent knows (or believes) is captured as set of sentences, whose content reflects what the agent knows (or believes). Alechina and Logan (2002), des Rivieres and Levesque (1686), Eberle (1974), Konolige (1986), Moore and Hendrix (1979), and Morreau and Kraus (1998) all take an approach along these lines. Philosophical motivation for their approach may be found in Quine, who sees propositional attitudes as 'involving something like quotation of one's own imagined verbal response to an imagined situation' (Quine 1960, 219).

Syntactic models of knowledge include a database $\mathscr{D}$ of formulas, and take KA to hold iff $A \in \mathscr{D}$. $\mathscr{D}$ is logically unstructured. It is merely a set of formulas, having no non-trivial logical closure features. As a result, all closure conditions for K are destroyed as well. But the solution seems cheap: knowledge or belief in syntactic structures have few interesting features. As Fagin et al. (1995) say,

> One gains very little intuition about knowledge from studying syntactic structures … In these approaches knowledge is a primitive construct. … Arguably, these approaches give us ways of *representing* knowledge rather than *modelling* knowledge. In contrast, the semantics given to knowledge in Kripke structures *explains* knowledge as truth in all possible worlds. (Fagin et al. 1995, 345)

An alternative approach draws on Scott-Montague *neighbourhood semantics* (Scott 1970), and represents what an agent knows or believes as an unstructured set $\mathscr{P}$ of propositions, rather than formulas. Propositions here are understood as sets of possible worlds. $\mathscr{P}$ is then an unstructured set of sets of worlds. This approach invalidates various forms of logical omniscience.

One that remains valid is (C4), though: if A and B are equivalent and KA holds, then KB holds as well. In a possible worlds setting,

that 2 + 2 = 4 and *that* $x^n + y^n = z^n$ *has no solutions in integers for* $n > 2$ are the same proposition, namely the total set of worlds. Yet one can know the former without knowing the latter.

Some proposals combine a syntactic and a possible worlds approach. *Awareness logic* (Fagin and Halpern 1988) works with three main notions. *Awareness* is syntactic: an agent is aware of *A* when *A* belongs to a set of formulas, its 'awareness set'. *Implicit knowledge* gets the standard possible worlds definition, whereas *explicit knowledge* is defined as the combination of implicit knowledge and awareness. An agent explicitly knows that *A* when she implicitly knows that *A* and *A* is in her awareness set. The underlying idea is that lack of omniscience can come from lack of awareness, understood as lack of conception (Schipper 2015). Because explicit belief requires awareness, and awareness is represented via membership of an arbitrary set of formulas, explicit belief can invalidate any non-trivial logical closure condition.

Differentiating between explicit and implicit representational mental states is as such cognitively plausible and independently motivated. The distinction between explicit and implicit memory and knowledge is frequently made in empirical psychology. Explicit memory is often taken to be conscious, involving the deliberate recall of previously acquired information. Implicit memory, by contrast, is taken to involve a change in performance or in the execution of a task in the light of previously acquired information without conscious recall (Schacter 1986, Schacter and Tulving 1994).

Similarly, representational accounts of belief claim that one believes *A* explicitly when one has a representation with content *A* actually present in the mind, as a token of a sentence inscribed in a 'belief box'. One believes *A* implicitly if one believes *A* without having a representation with that content present in one's mind (Dennett 1978, 1987). Dennett proposes that, in order for one to believe something implicitly, it is enough that 'the relevant content be swiftly derivable from something one explicitly believes' (Schwitzgebel 2015, §2.2.1).

What does *swift derivability* mean here? In the Fagin-Halpern approach, implicit belief is closed under classical logical consequence.

But this is precisely what spells trouble in the light of logical omniscience. In what sense does a finite cognitive agent have implicit belief in the infinitely many logical consequences of what it believes? This notion of implicit belief (or knowledge) seems more like the implicit rational commitments of one's beliefs (or knowledge). That's not the psychologically motivated notion of implicit belief just discussed (for criticisms along these lines, see (Schipper 2015, 88)).

But now, without a satisfactory notion of implicit knowledge or belief, the awareness approach fares little better than a purely syntactic approach to knowledge and belief representation (Jago 2006, Konolige 1986). Konolige sums up the situation:

> the logic of general awareness represents agents as perfect reasoners, restricted to considering some syntactic class of sentences. There don't seem to be any clear intuitions that this is the case for human or computer agents. (Konolige 1986, 248)

We now turn to approaches that resort to impossible worlds, with the aim of seeing whether some of them can do better.

5.3 Impossible Worlds for Knowledge and Belief

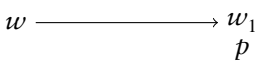
The idea of adopting impossible worlds in order to address the logical omniscience problem goes back to Hintikka (1975). He proposed that epistemically accessible worlds need not be genuinely logically possible. Instead, he allows the epistemically accessible worlds to be 'options which only look possible but which contain hidden contradictions' (Hintikka 1975, 476).

Rantala (1982a) gives an example of this strategy. Recall how, in Rescher and Brandom's work (§4.4), there are worlds at which conjunction and disjunction behave abnormally. In Rantala's approach, the idea is extended to all logical operators. Take once again the language $\mathscr{L}$ from §4.1 and give it the following semantics. A *Rantala frame* $\mathscr{F}$ for $\mathscr{L}$ is a triple $\langle\!\langle W, N, R \rangle\!\rangle$, with W the set of worlds,

$N \subseteq W$ the subset of normal, possible worlds, $W - N$ the non-normal or impossible worlds. R is as before. A frame becomes a Rantala model $\mathscr{M} = \langle\langle W, N, R, v\rangle\rangle$ when endowed with a valuation function v assigning truth values to formulas at worlds. At possible worlds in N, atomic formulas are directly assigned 1 or 0, and compound formulas are evaluated recursively. At impossible worlds in $W - N$, by contrast, *all* formulas are assigned a truth value by v directly, not recursively. Logical consequence and validity are defined, again, as truth preservation at all possible worlds in all models.

As a consequence, at impossible worlds, all formulas are treated as if they are atomic. $A \vee B$ may turn out to be true even though both A and B are false (impossible worlds may be non-prime), and $\neg A$ may turn out to be true when A is (impossible worlds may be inconsistent). This is the generalization of a strategy already met in §4.2: taking impossible worlds as worlds where the logical syntax of formulas can be disregarded when assigning them a truth value. As a consequence, the impossible worlds in $W - N$ are not closed under any consequence relation other than *identity*, $A \vDash A$. Because of this, impossible worlds of this kind are called *open worlds* in Priest (2005).

On this approach, none of of (C1)–(C8) hold. For instance, against (C1) and (C8), consider the following model, with $N = \{w\}$ and the arrow representing accessibility:



At w, Kp is true (since p is true at w_1) but $K(p \vee q)$ is not (since $p \vee q$ is false at w_1). So $Kp \nvDash K(p \vee q)$, even though $p \vDash p \vee q$. (Note that this is no countermodel to $p \vDash p \vee q$. Since $w_1 \in W - N$, it does not affect logical consequence.)

Rantala (1982b) extends this approach to quantified modal logics, and Wansing (1990) develops it into a unified framework for epistemic logics. Wansing shows that various logics for knowledge and belief in Artificial Intelligence, including Fagin and Halpern's

awareness logic (mentioned in §5.2), have equivalent impossible worlds models (that is, models which validate precisely the same formulas). Sillari (2008) establishes further equivalence results in the area.

A naive Rantalian approach, however, faces a serious problem. As noted by Jago (2007, 2009b), the way in which Rantala models manage to invalidate all forms of logical omniscience involves having no restriction on the impossible worlds one can look at, via the epistemic accessibility relation R of the models. The set of worlds W in the frames can include worlds not closed under any non-trivial consequence relation. Worlds, thus, can correspond to arbitrary sets of formulas of $\mathscr{L}$.

Given world w where our epistemic agent is located, then, let $\mathcal{S}_w = \{w_1 \mid Rww_1\}$, the set of worlds accessible from w. Let $\mathscr{C} = \{A \mid v_{w_1}(A) = 1 \text{ for all } w_1 \in \mathcal{S}_w\}$, the set of formulas true at all of them. The agent's epistemic or doxastic state can be reduced to a merely syntactic structure: KA holds at w just in case $A \in \mathscr{C}$, and $\mathscr{C}$ can be a set of formulas lacking any (non-trivial) closure property. The content of epistemic states, then, comes out as highly structured as the syntax of the language. 'Unconstrained' impossible worlds semantics makes no real progress with respect to a merely syntactic approach. One can add constraints on the accessibility relation, or on the logical behaviour of the accessible worlds, which will validate some inferences. As we are about to see, though, how this should be done is no trivial issue.

5.4 Closure under a Weaker Logic

We've seen that, by adding impossible worlds with no logical structure whatsoever, the worlds approach seems no better than the syntactic approach. A natural thought is that we should insist that impossible worlds have *some* degree of logical structure, although not as much as logically possible worlds. Cresswell (1973) and Levesque (1984)

offer approaches along these lines. We'll present a simple version in this section.

Our language $\mathscr{L}$ will include the operators $\neg$, $\wedge$, $\vee$, and K, with $A \supset B$ being defined as $\neg A \vee B$. As before, a frame $\mathscr{F}$ is a pair $\langle\langle W, R\rangle\rangle$, with W the set of worlds and R the epistemic accessibility relation. A frame becomes an *FDE model* $\mathscr{M} = \langle\langle W, R, \rho\rangle\rangle$ when endowed with a *valuation relation* ρ. (We'll explain the name 'FDE' below.) Unlike the usual valuation function, a valuation relation can connect a formula to more than one truth value at a world: ρ can relate the atomic formulas of $\mathscr{L}$ to truth ($\rho_w p1$), falsity, ($\rho_w p0$), both, or neither. We thus get rid of the assumption, embedded in the semantics of classical logic, that truth and falsity are exclusive and exhaustive.

We extend ρ to the whole language via the following recursive clauses. We now need to spell out truth and falsity conditions separately for each operator, for now not being true (not being related to 1) is distinct from being false (being related to 0):

(S1$\neg$) $\rho_w(\neg A)1$ iff $\rho_w A0$

(S2$\neg$) $\rho_w(\neg A)0$ iff $\rho_w A1$

(S1$\wedge$) $\rho_w(A \wedge B)1$ iff $\rho_w A1$ and $\rho_w A1$

(S2$\wedge$) $\rho_w(A \wedge B)0$ iff $\rho_w A0$ or $\rho_w B0$

(S1$\vee$) $\rho_w(A \vee B)1$ iff $\rho_w A1$ or $\rho_w B1$

(S2$\vee$) $\rho_w(A \vee B)0$ iff $\rho_w A0$ and $\rho_w B0$

(S1K) $\rho_w(KA)1$ iff for all $w_1 \in W$ such that Rww_1, $\rho_{w_1} A1$

(S2K) $\rho_w(KA)0$ iff it is not the case that $\rho_w(KA)1$

Logical consequence is truth preservation at all worlds of all models:

> $\Gamma \vDash A$ iff for all models $\mathscr{M} = \langle\langle W, R, \rho\rangle\rangle$ and all $w \in W$: if $\rho_w B1$ for all $B \in \Gamma$, then $\rho_w A1$

(Note that we define logical consequence over all worlds in the model, making no distinction between possible and impossible worlds here.)

This K operator corresponds to what Levesque (1984) calls *explicit* belief. He also defines an implicit belief operator, which is closed under classical logical consequence and hence delivers full logical omniscience (for the notion of implicit belief). We'll focus on the explicit notion only here.

In this semantics, worlds can be inconsistent (making both A and $\neg A$ true, for some A) and incomplete (making neither A nor $\neg A$ true). In the taxonomy of §1.4, they are impossible worlds of the third and fourth kinds: violating classical logic and making contradictions true. Yet they do have some logical structure. They are still *adjunctive*, making $A \wedge B$ true whenever they make both A and B true, and *prime*, making either A or B true whenever they make $A \vee B$ true. They also obey Disjunction Introduction (from A to $A \vee B$) and Double Negation Introduction and Elimination (from A to $\neg\neg A$ and back).

The resulting logic is *paraconsistent*, for $A \wedge \neg A \not\models B$: contradictions do not entail arbitrary conclusions. It is also *paracomplete*, for $A \not\models B \vee \neg B$: arbitrary premises do not entail all instances of Excluded Middle. The extensional fragment of this logic (the part lacking K-sentences) is, in fact, one way of presenting *First Degree Entailment*, **FDE** (Belnap 1977, Dunn 1976). This is a simple and well-known paraconsistent and paracomplete logic, and is why we call these structures *FDE models*. We'll also refer to the worlds in those models as *FDE worlds*.

FDE models avoid some problematic forms of logical omniscience. The approach can be used to model agents that have contradictory beliefs (against (C3)), but do not thereby believe everything. Consider the model:

(We display all the atoms related to a truth-value by ρ.) Then $K(p \wedge \neg p)$ holds at w, but Kq does not.

Closure under known (or believed) implication (C6) fails too. This is because *modus ponens* fails in the semantics: $A, A \supset B \not\models B$. In the model above, *modus ponens* fails at w_1: $p \supset q$ holds there, for any q, given that $\rho_{w_1} p0$ and that $p \supset q$ is defined as $\neg p \vee q$. For any world that can access such worlds, it may be that KA and $K(A \supset B)$ both hold, and yet KB does not. So in general, $KA, K(A \supset B) \not\models KB$.

Fagin and Halpern (1988) and Jago (2007) stress that logical omniscience strikes back in unwelcome ways, however. Any epistemically accessible FDE world will be adjunctive, closed under Disjunction Introduction, and Double Negation Introduction and Elimination. As a consequence, knowledge and belief come out closed under the corresponding entailments. Conditions (C7) and (C8) still hold.

It is questionable whether this is a good way to model finite and fallible epistemic agents. In particular, one may believe that A without believing that $A \vee B$ for an arbitrary B, for one may lack the B-involving concepts. Similarly, any FDE world making A true also makes true the formula obtained by prefixing an even number of negations to A. But it might be that an agent believes that A, without believing that

$$\overbrace{\neg\neg \cdots \neg\neg}^{\text{1,000,000 times}} A$$

simply because she lacks the cognitive resources to record all the iterations.

More generally, in this system knowledge and belief are closed under weaker-than-classical first-degree entailment. If KA and B is an **FDE**-consequence of A, then KB. But this seems wrong: finite cognitive agents do not know or believe all remote consequences of what they know or believe, even when the notion of entailment in play is **FDE**. There are infinitely many such consequences and they cannot all be computed by a finite mind.

We seem to face a dilemma: either cognitive states like knowledge and belief for real agents are closed under some logical consequence, or they're not. If they are, then logical omniscience returns: we have the implausible situation of an agent that is omniscient with respect

to the target logic. For this not to happen, we seem to have to admit that such states are completely anarchic: they violate any logical closure principle (except for A entailing A). But then, how can we have a *logic* of knowledge and belief at all? Some logicians and AI researchers have conjectured that there is no solution to the dilemma (Meyer and Van der Hoek 1995, 88).

5.5 Going Dynamic

We will now focus on a version of this dilemma, phrased directly in terms of worlds, which we take from Bjerring (2010, 2012). (See also Jago 2014a.) Take a set $\mathscr{R}$ of rules of inference. One may think of a set of sequent calculi or natural deduction rules; but (bracketing issues caused by the presence of rules that discharge temporary assumptions) the point is independent from the specific logical set-up. Call an inference from $A_1, \ldots, A_n$ to B *immediate* when it involves just a single application of a single rule in $\mathscr{R}$.

If what one believes is represented by a set of accessible worlds and these are all closed under the rules in $\mathscr{R}$, then one turns out to be omniscient with respect to $\mathscr{R}$. Suppose, instead, that some accessible world w is not closed under some rule $r \in \mathscr{R}$. Then for some B that follows immediately from some $A_1, \ldots, A_n$, w makes true all of $A_1, \ldots, A_n$ but not B. Then our agent is represented as missing some immediate consequence of what she believes. She considers w as a way things might be, even though it is, in an *obvious* way, not a way things could be. This seems to be a poor approach to modelling our agent's rational states.

The difficulty we have in deciding which impossible worlds may be accessible to our non-idealized agents is due to the fact that, seemingly, there is no third option: either epistemically accessible worlds are closed under full logical consequence, or they turn out to be obviously impossible. That, in short, is what Jago (2014a) calls *Bjerring's Problem* (which we'll discuss in more detail in §10.3).

What seems especially difficult to do is to model epistemic agents that are rationally *competent*, in spite of not being omniscient. Real agents just cannot believe all that follows from what they believe. But if 'anything goes', so that believing something does not entail believing anything else in particular, then it seems that we are modelling agents who are not even moderately rationally competent: they fail to believe obvious consequences of what they believe.

One recent approach to the issue develops an idea for modelling competent but non-omniscient agents dynamically, in terms of how their beliefs will or may evolve over time due to epistemic actions and events. As a response to the problem of modelling competent but non-omniscient agents, the idea was originally put forward by Duc (1995, 1997). Bjerring and Rasmussen (2018) and Rasmussen (2015) update the approach, using the *dynamic epistemic logics* framework. To evaluate their approach, we'll need to understand a little about dynamic epistemic logic.

Dynamic logics contain operators based on actions. If a is an action, then '$[a]$' is an operator, and '$[a]A$' says that, after action a has been carried out, A is the case. Semantics for these operators is given in terms of transformations on the model. Typically, the semantics of such logics uses pointed Kripke models, that is, models $\mathscr{M}^w$ with a regular Kripke model $\mathscr{M}$ and a 'designated' base world w, which is itself in $\mathscr{M}$ and can be thought of as the world considered to be the actual one. Then $[a]A$ is true in $\mathscr{M}^w$ iff A is true in all pointed models $\mathscr{N}^{w_1}$ obtained by transforming $\mathscr{M}^w$ according to the instructions encoded in action a. Dynamic epistemic logic, developed by Baltag et al. (1998), Segerberg (1995), Van Benthem (2011), Van Ditmarsch (2005), and Van Ditmarsch et al. (2008), adds epistemic operators to the dynamic semantics. Baltag and Renne (2016) give an introduction to the approach.

Bjerring and Rasmussen (2018), following ideas in Rasmussen 2015, adapt this approach to model competent but non-omniscient agents. In their models, agents count as competent insofar as they unfold the consequences of their beliefs, up to a certain 'depth' of reasoning. Their key dynamic operator is of the form '$\langle n \rangle A$', to be

read as: 'After n steps of logical reasoning, A may be the case'. These steps of reasoning are n applications of rules from a chosen set $\mathscr{R}$. The approach also has an epistemic operator B, allowing for sentences of the form '$\langle n\rangle BA$', saying that the agent can come to believe that A after n steps of reasoning.

Models (adapted to our own notation in this book) are tuples $\mathscr{M} = \langle\!\langle W, N, f, v\rangle\!\rangle$, where W is the total set of worlds, $N \subseteq W$ is the set of normal-possible worlds, f is an epistemic projection function (see §5.1) mapping each world to the set of worlds epistemically accessible from it, and v is a valuation function. Pointed models are model-world pairs, written $\mathscr{M}^w$, where $w \in N$.

Given a pointed model $\mathscr{M}^w$ and a set of rules $\mathscr{R}$, we can define a set of epistemic projection functions $\mathscr{F}^n$ for each integer n. Intuitively, these functions capture all chains of reasoning of length n using rules in $\mathscr{R}$, thus, all ways in which the agent can modify the set of worlds initially seen as epistemically possible, by performing a chain of reasoning steps of length n. We then define an equivalence relation $\sim^n$ for each n, relating pointed models that differ at most in their projection functions f, which must be chosen from $\mathscr{F}^n$. (We skip the definitions here: see Bjerring and Rasmussen (2018) for the details. We'll present a related rule-based approach in detail in §10.5.)

We then extend each pointed model $\mathscr{M}^w$'s valuation function v to all formulas. The clauses for connectives and B are as in standard epistemic logic, and the clause for $\langle n\rangle A$ is given as follows (at $w \in N$; impossible worlds have complex formulas evaluated directly, non-recursively):

(S$\langle n\rangle$) $v_w(\langle n\rangle A) = 1$ iff there is some pointed model $\mathscr{N}^w$ with extended valuation v' such that $\mathscr{M}^w \sim^n \mathscr{N}^w$ and $v'_w A = 1$.

The Bjerring-Rasmussen framework is a plausible and promising model of how non-omniscient agents can deductively unfold the consequences of their beliefs, and update their belief states accordingly. For, given a set of inference rules, we can align what the agent *can* come to believe in n steps of reasoning with possible proofs containing

no more than n steps. Bjerring and Rasmussen then prove that, if C follows from $A_1, \ldots, A_m$ in n steps of reasoning given rules $\mathscr{R}$, then $BA_1, \ldots, BA_m$ together entail $\langle n \rangle BC$ (Bjerring and Rasmussen 2018, 17, Corollary 1).

Their overall aim, however, is to capture a non-omniscient agent's logical competence, for which they provide the following behavioural test:

> For any p and q such that q follows trivially from p, if an agent believes p, then upon being asked whether q is the case does she immediately answer 'yes'? If she does, she passes the test and counts as moderately logically competent. (Bjerring and Rasmussen 2018, 3)

One might quibble with this test (after all, agents can be confused about what they believe; they often speak insincerely; and often require a moment's reflection before asserting). But let's accept it for our evaluation of the approach.

Does Bjerring and Rasmussen's approach pass the test they set? This is formulated in terms of what an agent will do: she will answer 'yes' when asked whether q, a trivial consequence of her beliefs. But the formal approach tells us about what an agent *can* come to believe, within n steps of reasoning. The main result is that $BA_1, \ldots, BA_m$ together entail $\langle n \rangle BC$ when C follows from the A_is within n steps. $\langle n \rangle BC$ says that the agent can come to believe C within n steps: there is some n-step chain of reasoning the agent can follow, via which she will come to believe that C. This isn't the same as telling us that our agent will come to believe that C. For she might follow some other chain of reasoning, and thereby come to believe something other than C. So there's no guarantee that the modelled agent will answer 'yes' when asked whether q. As such, that agent hasn't been show to pass the test for moderate logically competence.

If we're interested only in agents which pass Bjerring and Rasmussen's test, then we need to focus on what the modelled agent *will* come to believe after n steps, however she reasons. We need to switch from particular to universal quantification over pointed models, that

is, to a box-like dynamic modality, '$[n]$'. $[n]A$ is true on a pointed model when A is true on all $\sim^n$-related pointed models. (Then $[n]A$ is equivalent to $\neg\langle n\rangle\neg A$. But we can't define $[n]A$ directly in this way, since $[n]A$ and $\neg\langle n\rangle\neg A$ may come apart at impossible worlds.)

It's easy to see that, if the rules in $\mathscr{R}$ are sufficiently general, then there's nothing the modelled agent *must* believe after n steps (for any finite n). Suppose $\mathscr{R}$ contains Disjunction Introduction and our agent believes that p. Then one chain of reasoning goes: p, $p \vee p$, $p \vee p \vee p$, and so on. Another goes: p, $p \vee q$, $p \vee q \vee q$, and so on. The agent can go for either chain of reasoning. So the only thing she is guaranteed to believe after n steps of reasoning is p, which she believed to begin with. In other words, $[n]BA$ entails BA.

It also turns out that the move from one model to a $\sim^n$-related one is belief-monotonic, which means that an agent's current beliefs are preserved (with additions, but no subtractions) in the new model. This guarantees that BA entails $[n]BA$, and hence that $[n]BA$ is equivalent to BA. So formulas of the form '$[n]BA$' tell us nothing about logical competence, over and above what formulas of the form 'BA' tell us. But then, the modelled agent will fail the test for moderate logical competence, since beliefs aren't closed under any notion of consequence. Neither dynamic modality, then, helps us capture the target concept of moderate logical competence.

Bjerring and Rasmussen's behavioural test is (rightly, in our opinion) a normative one. It sets up a standard to be met, and so tells us something about what's expected of agents. Ordinary reasoners can fail the test, of course. We all make mistakes in our reasoning from time to time, and thereby, on those occasions, fail to live up to rational standards. In a normative setting, the aim is not to model what an agent can come to believe. Rather, it's to model a normative notion of belief, which builds in a certain amount of rationality, without idealizing agents to the point of logical omniscience.

Much more needs to be said on the key ideas we've introduced in this section. One is that of a *trivial inference*. We offer an analysis in §9.5. Another is the idea that epistemically accessible worlds cannot be obviously impossible. This leads to Bjerring's Problem: that

worlds are either closed under full logical consequence, or obviously impossible, and neither should be epistemically accessible. We offer a philosophical analysis of the problem, and attempt to draw a distinction between *obvious* and *subtle* impossibilities within a formal model, in §9.5. Finally, there is the normative, rational, but non-ideal notion of belief we've just discussed. We propose a model of the notion in §10.5, based on the idea of subtly impossible worlds from §9.5.

Chapter Summary

Standard possible-worlds epistemic logic gives rise to the problem of logical omniscience (§5.1). There are attempts to deal with the problem without using impossible worlds. We discussed a number of these approaches, and found them all wanting (§5.2). The impossible worlds approach is immediately more successful, but faces a deep problem: how should impossible worlds be constrained, so as to give adequate models of knowledge and belief (§5.3)? One option is to take impossible worlds to be closed under some weaker-than-classical logic. But this approach does not genuinely solve the problem of logical omniscience (§5.4).

A different approach is the *dynamic* one, whereby epistemic states are not closed at any one time, but nevertheless evolve towards closure in a dynamic way (§5.5). We found this approach promising, but we will propose an alternative philosophical account of epistemic and doxastic states in §10.5.

6

Relevant Logics

6.1 Basic Relevant Logic

Relevant logic (or, as they call it in the US, *relevance* logic) aims at developing a notion of conditionality free from the 'fallacies of relevance': the 'paradoxes' of the material and strict conditional (§1.3), including:

(6.1) $A \rightarrow (B \rightarrow B)$

(6.2) $A \rightarrow (B \vee \neg B)$

(6.3) $(A \wedge \neg A) \rightarrow B$

Whether '$\rightarrow$' is understood as material or strict, such conditionals come out logically valid in ordinary modal logic with possible worlds semantics. However, in each case, there need be no real connection between antecedent and consequent. 'If Amsterdam is in the Netherlands, then if snow is white, then snow is white' is an instance of (6.1), yet what has Amsterdam's being in the Netherlands to do with (trivial consequences of) snow's being white?

We can debate whether (6.1)–(6.3) are *really* invalid. 'Be relevant' was one of Grice's conversational maxims, the *Maxim of Relation*. Perhaps what's wrong with (6.1)–(6.3) is that they violate this pragmatic rule. Perhaps that's all that's wrong with them. They might be logically fine after all, but pragmatically difficult to assert. For relevant logicians, however, relevance is not merely pragmatics: it is

an integral feature of logic and formal semantics. They argue that this approach is more plausible than starting with classical logic, and then claiming that irrelevancies should not be asserted purely on the basis of Gricean maxims.

Anderson and Belnap (1975) pioneered relevant logics with the aim of avoiding irrelevance in logic. They held that a formula of the form $A \rightarrow B$ should be a theorem only if A and B shared a sentential atom or parameter. This was called the *Variable Sharing Property* (Dunn and Restall 2002, 27), and was meant to capture the idea of a real connection between antecedent and consequent. They intended the Variable Sharing Property to be necessary, but not sufficient, for a conditional to count as being relevantly acceptable.

Anderson and Belnap initially proceeded proof-theoretically, writing down lists of axioms and rules of inference which gave to the conditionals of their systems the required features. Soon, however, the issue of providing a semantics for such systems came to the top of the agenda:

> Yea, every year or so Anderson & Belnap turned out a new logic, and they did call it E, or R, or E_I, or $P - W$, and they beheld such logic, and they were called relevant. And these logics were looked upon with favor by many, for they captureth the intuitions, but by many they were scorned, in that they hadeth no semantics. Word that Anderson & Belnap had made a logic without semantics leaked out. Some thought it wondrous and rejoiced, that the One True Logic should make its appearance among us in the Form of Pure Syntax, unencumbered by all that set-theoretical garbage. Others said that relevant logics were Mere Syntax. (Routley and Meyer 1973, 194)

Anderson and Belnap (1975, §28.2) later found algebraic semantics, with soundness and completeness proofs, for these logics, or fragments of them. Usually, these made use of algebraic structures called *De Morgan lattices*. But in such approaches, the syntax and the semantics seem to copy each other. Lacking an independent understanding of the latter, these results often leave philosophically inclined logicians unsatisfied.

In other words, Anderson and Belnap's approach looks like *pure*, rather than *applied*, semantics. (The terminology may be due to Plantinga (1974), although Carnap (1948) and Dummett (1973) make similar distinctions.) Pure formal semantics consists in mathematical structures which interpret the language but which have a merely mathematical meaning. In applied formal semantics, by contrast, we have a clear understanding of the connection between the mathematics and meaning. For Dummett, the former are of 'purely technical' interest, whereas the latter 'are taken to have a direct relation to the use which is made of the sentences of a language' (Dummett 1973, 6–7).

Relevant logicians sought to develop frame semantics for relevant logics which promised to move beyond syntax or pure algebra. But this seemed difficult. A logical truth is true everywhere; so how could a conditional with it as a consequent fail to be true everywhere? This requires a semantics that can, at the same time, (a) account for failures of logical truths, but (b) not relinquish their status as logical truths. The 'Routley-Meyer semantics' of Routley and Meyer 1973 was a breakthrough. They distinguished two kinds of points, *normal* and *non-normal*, and this allowed them to invalidate conditionals like (6.1)–(6.3). Non-normal points are naturally interpreted as impossible worlds (Priest 2008, 171–4).

Let us introduce a simple frame semantics of this kind. (We now follow Restall 1993, a simplified version of the original Routley-Meyer semantics presented also in Priest 2008, chapter 10.) $\mathscr{L}$ is as before but with a conditional $\rightarrow$. A *Routley-Meyer frame* $\mathscr{F}$ for $\mathscr{L}$ is a quadruple $\langle\!\langle W, N, R, * \rangle\!\rangle$, with W the set of worlds, $N \subseteq W$ the subset of normal worlds, and $W - N$ the non-normal worlds. $R \subseteq W \times W \times W$ is a ternary relation on worlds satisfying a *Normality Condition*:

(NC) If $w \in N$, then Rww_1w_2 iff $w_1 = w_2$.

Finally, $*$ is the *Routley Star*: a *period two* operation on W ($w^{**} = w$ for each $w \in W$). Given w, let us call w^* the *twin* of w. We will come to the issue of what R and $*$ may mean later on.

A frame becomes a model $\mathscr{M} = \langle\langle W, N, R, *, v\rangle\rangle$ when endowed with a valuation function v, assigning truth values to AT at worlds. We extend v to the whole language as follows:

(S¬) $v_w(\neg A) = 1$ if $v_{w^*}(A) = 0$, and 0 otherwise.

(S∧) $v_w(A \wedge B) = 1$ if $v_w(A) = v_w(B) = 1$, and 0 otherwise.

(S∨) $v_w(A \vee B) = 1$ if $v_w(A) = 1$ or $v_w(B) = 1$, and 0 otherwise.

(S→) $v_w(A \to B) = 1$ if for all $w_1, w_2 \in W$ such that Rww_1w_2, if $v_{w_1}(A) = 1$ then $v_{w_2}(B) = 1$, and 0 otherwise.

Logical validity and consequence are truth/truth preservation at all normal worlds in all models. The logic which is sound and complete with respect to this semantics is called **B**, for *Basic relevant logic*. (Don't confuse this with the modal logic from §4.1, also called **B**.)

The effect of (NC) is that $A \to A$ holds at all normal worlds. For $A \to A$ to hold at w, we require that, if Rww_1w_2 and A holds at w_1, then A holds at w_2 also. But if w is a normal world, then $w_1 = w_2$, and hence A holds at w_2 by assumption. Since validity is defined as truth at all normal worlds, it follows that $A \to A$ is valid in **B**. But it does not follow that each instance of $A \to A$ holds at all worlds. For now suppose w_1 is a non-normal world, and that $Rw_1w_2w_3$, where $w_2 \neq w_3$ and q holds at w_2 but not at w_3. Then $q \to q$ does not hold at w_1.

Now let's see how this allows us to deal with (6.1), $A \to (B \to B)$. We need a world at which A holds but $B \to B$ fails. We can use the model from above, in which $q \to q$ fails at w_1. Now add to the model a normal world w such that Rww_1w_1, and add that p holds at w_1. Then by definition, $p \to (q \to q)$ fails at w. Since w is a normal world, what holds there determines what's valid. Since an instance of $A \to (B \to B)$ fails at w, it isn't valid.

To understand how (6.2) and (6.3) are handled, we need to say something about the Routley Star. The clause (S¬) has it that $\neg A$ is true at a world w if and only if A is false at its twin, w^*. When w and w^* are distinct, we cannot read the value of $\neg A$ at w from the value of

A at w (as we could if $\neg$ were classical negation). Relevant negation is a modal operator: in order to evaluate negated formulas at w, one has to look at the goings on of a world that may be distinct from w. This negation is often called *De Morgan negation* in the literature, for all of De Morgan's Laws hold for it:

$$\neg\neg A \dashv\vDash A \qquad \neg(A \wedge B) \dashv\vDash \neg A \vee \neg B \qquad \neg(A \vee B) \dashv\vDash \neg A \wedge \neg B$$

This set-up delivers worlds which are classically impossible, for Excluded Middle fails at them; and contradiction-realizers where A and $\neg A$ are both true. This is what we need to have (6.2), $A \rightarrow (B \vee \neg B)$, and (6.3), $(A \wedge \neg A) \rightarrow B$, fail. For (6.2), a counterexample is given by a normal world w such that Rww_1w_1, with w_1 a world (which may be normal, or not) where p is true but neither q nor $\neg q$ are. This happens, by (S$\neg$), when w_1 is such that q is false at it but true at its twin, w_1^*. For (6.3), a counterexample is given by a normal world w such that Rww_1w_1, with w_1 a world (normal or not) where both p and $\neg p$ are true but q is not. This happens, by (S$\neg$) again, when w_1 is such that q is true at it but false at w_1^*.

One may not like the idea of normal worlds – which we think of as possible worlds – where contradictions are true, or where Excluded Middle fails. To alleviate such worries, one can add the *Classicality Condition* as a constraint on the semantics:

(CC) If $w \in N$, then $w = w^*$.

When a world w is its own twin, then by (S$\neg$) the behaviour of negation at it is just the classical one: $\neg A$ is true at w just in case A is false at the very same world. That world is maximally consistent, making true exactly one of A and its negation. (CC) guarantees that all normal worlds are like that. This gives a logic stronger than the basic relevant **B** but, with the proviso that w_1 is a non-normal world, the counterexamples to (6.1)–(6.3) still go through.

There is a certain translatability between the Routley Star semantics for negation and the relational semantics for negation we met in §5.4. Suppose we define $\rho_w p1$ if and only if $v_w p = 1$, and $\rho_w p0$ if and only

if $v_{w^*}p = 0$. Then ρ will work just like a pair consisting of a w and its twin w^* (Priest 2008, 153). In fact, these are equivalent ways of presenting negation in **FDE** (§5.4). One reason for resorting to the Routley Star in this chapter is that adding a relevant conditional using the ternary R to the relational semantics does not give, in any but the simplest cases, the usual family of relevant logics.

On the Routley-Meyer semantics, the corresponding logics are, in an important way, *automatically* relevant. The Variable Sharing Property falls straight out of the Routley-Meyer semantics. As Priest says of this approach,

> relevance is not some extra constraint imposed on top of classical validity. Rather, relevance, in the form of parameter sharing, falls out of something more fundamental, namely the taking into account of a suitably wide range of situations. (Priest 2008, 174)

Contrast this situation with *filter logics* (Smiley 1959, Tennant 1984), in which relevant validity for conditionals is classical validity plus a constraint that filters out irrelevancies. There's something *ad hoc* about that approach. The Routley-Meyer semantics, by contrast, seems to be a natural extension of frame semantics with a binary relation R, from which the Variable Sharing Property naturally falls. What's crucial to this approach is that Priest's 'range of situations' is wider-than-classical, in that it now comprises non-normal, impossible worlds. These are essential to the frame semantics of mainstream relevant logic.

6.2 Stronger Relevant Logics

We can impose further constraints on the ternary relation R to obtain stronger relevant logics, which validate more principles than those of **B**. (This is similar to the way we obtained modal logics stronger than **K** in §4.1 by imposing conditions on the binary relation R.) The following conditions:

(6.4) If Rww_1w_2, then $Rww_2^*w_1^*$

(6.5) If there is a $x \in W$ such that Rw_1w_2x and Rxw_3w_4, then there is a $y \in W$ such that Rw_1w_3y and Rw_2yw_4

(6.6) If there is an $x \in W$ such that Rw_1w_2x and Rxw_3w_4, then there is a $y \in W$ such that Rw_2w_3y and Rw_1yw_4

validate, respectively, the following principles:

(Contraposition) $(A \to \neg B) \to (B \to \neg A)$

(Suffixing) $(A \to B) \to ((B \to C) \to (A \to C))$

(Prefixing) $(A \to B) \to ((C \to A) \to (C \to B))$

These are all desirable principles. With all three accepted, the corresponding logic is often called **TW**.

To add further desirable principles, we can add extra information to the Routley-Meyer frames, in the form a binary relation $\preccurlyeq$ between worlds. An *expanded Routley-Meyer frame* $\mathscr{E}$ is now a quintuple $\langle\!\langle W, N, R, *, \preccurlyeq \rangle\!\rangle$, which becomes an expanded model when the interpretation function v is added. All the familiar components are as before, and $\preccurlyeq \subseteq W \times W$ is a binary relation between worlds satisfying the following conditions. If $w \preccurlyeq w_1$, then:

(6.7) If $v_w(p) = 1$, then $v_{w_1}(p) = 1$

(6.8) $w_1^* \preccurlyeq w^*$

(6.9) If $Rw_1w_2w_3$, then ($w \in N$ and $w_2 \preccurlyeq w_3$) or ($w \notin N$ and Rww_2w_3)

Think of '$w \preccurlyeq w_1$' as saying that w_1 inherits the truths of w. (6.7) says that this is so for atomic truths, and (6.8) and (6.9) generalize to all formulas. Together, these guarantee that, if $w \preccurlyeq w_1$, then $v_w(A) = 1$ only if $v_{w_1}(A) = 1$, for all $A \in \mathscr{L}$.

There are two further important conditions, the second crucially involving $\preccurlyeq$:

(6.10) If $Rw_1w_2w_3$ then there is an $x \in W$ such that Rw_1w_2x and Rxw_2w_3

(6.11) If $Rw_1w_2w_3$ then there is an $x \in W$ such that $w_1 \preccurlyeq x$ and Rw_2xw_3

These validate, respectively:

(Contraction) $(A \to (A \to B)) \to (A \to B)$

(Assertion) $(A \to ((A \to B) \to B)$

Adding all of these principles give us the relevant logic **R**, possibly the best-known relevant system. It can be shown that the Variable Sharing Property holds in **R**, and hence in the weaker systems as well (see Priest 2008, 205–6).

Let us now move on to the main topic of this chapter: how to make sense of the mainstream Routley-Meyer semantics for relevant logics and, in particular, of its non-normal worlds.

6.3 Relevant Worlds as Information States

Giving an intuitively plausible reading of the ternary relation and of the Routley Star has proved difficult. Copeland (1979) and Van Benthem (1979) claimed that the Routley-Meyer frame semantics are pure, not applied, semantics, for no independent understanding had been offered of what the ternary R and Routley Star $*$ mean. Relevant logicians came to the rescue in various ways. We will now go through three main strategies, paying special attention to how the worlds of relevant frames are understood in them.

One well-known approach interprets worlds in relevant frames as states of information, and R as an informational relation on those states. Urquhart (1972) proposed that 'Rww_1w_2' be read as claiming that the information in w_2 is obtained by merging together the information in w and that in w_1.

If this information merging is understood as entailing that w_1-information is included in w_2-information, then the following would seem to hold:

(6.12) If Rww_1w_2, then $w_1 \preccurlyeq w_2$

But this principle makes $B \rightarrow B$ true at all worlds, non-normal as well as normal, which in its turn validates the irrelevant principle (6.1), $A \rightarrow (B \rightarrow B)$. So some other notion of information merging is required.

Systematic information-theoretic readings of the semantics have been proposed by Mares (2004) and Restall (1995). These rely on interpreting the worlds in the frames as representing *situations* in the sense of Barwise and Perry's *situation theory* (Barwise and Perry 1983).

In situation theory, situations are information-supporting structures, allowing the fine-grained distinctions unavailable within possible world semantics. Situations need not be maximal: they can fail to support information about certain topics. The situation consisting of Mark's office in Nottingham does not support the information that it is raining in Amsterdam, nor the information that it is not raining there. Situations may be abstract as well as concrete, and may represent logical impossibilities. Barwise and Seligman (1997) develop situation theory into a general theory of information flow in distributed systems.

Barwise and Parry's original approach did not rule out the possibility that situations act, not only as sites of information, but also as information channels or conduits between other situations. Restall (1995) adopts this view. The points of the relevant frame semantics are taken here as playing both roles. So the situation consisting of a living room with a TV turned on in Nottingham can support the information that it is raining in Amsterdam. It does so by connecting the rainy Dutch situation to the living room via the channel consisting of the cameras, wires, signals, and so on, connecting the two sites. We should read 'Rww_1w_2', then, as 'w is a conduit of information

from site w_1 to site w_2', or as 'situation w allows information to flow from situation w_1 to situation w_2'.

This helps to understand the semantic clause (S→) from §6.1. When w allows the information that $A \to B$ to flow from w_1 to w_2, and w_1 supports the information that A, then w_2 supports the information that B. Vice versa, if w does not allow the information that $A \to B$ to flow, there must be situations w_1 and w_2 such that w_1 supports the information that A, but w_2 does not support the information that B.

Mares (2004, 2009, 2010) too understands the worlds in relevant frames as information-conveying situations, in the sense of situation theory:

> a situation w can be said to contain the information that $A \to B$ if on the hypothesis that there is a w_1 in the same world that contains A, we can derive that there is a situation w_2 in the same world in which B. … [This theory] is about making inferences from the perspective of situations about the situations in a world. (Mares 2010, 211, notation modified)

(In Mares's terminology, there are both situations and worlds, and situations are included in worlds.) This analysis interprets R in terms of what can be derived from what, so that Rww_1w_2 says 'all the information that we can derive really using the information in both w and w_1 is all contained in w_2' (Mares 2010, 211, notation modified).

Mares thinks of this in terms of *situated inference*, facilitated by 'informational links' Mares (2004). An informational link is a 'perfectly reliable connection, such as a law of nature or a convention' (Mares 2009, 426). Similarly, Devlin (1991, 12) characterizes these constraints as 'natural laws, conventions, analytic rules, linguistic rules, empirical law-like correspondences, or whatever'. A sufficient condition for Rww_1w_2, for example, is that a law of nature of w relates w_1 to w_2. These informational links 'are themselves contained as information in situations' (Mares 2004, 44).

Mares (2004) also analyses *propositions* as sets of situations which satisfy some closure features. There is a relation of *situated implication*, Iww_1P, holding between situations w and w_1 and a proposition P. It holds when the information jointly supported by w

and w_1 allows us to infer the existence of a situation where P holds. We then read 'Rww_1w_2' as 'w_2 belongs to every proposition P such that Iww_1P'. When w supports the informational link $A \rightarrow B$, and w_1 supports the information that A, then w_2 represents a situation belonging to the proposition expressed by B. Vice versa, if w does not support $A \rightarrow B$ there must be some w_1 and w_2 such that w_1 supports the information that A, while w_2 does not belong to the proposition expressed by B.

Although these interpretations focus first of all on making sense of the ternary R, both Restall (1999) and Mares (2004) offer an interpretation of the Routley Star, too. They use a binary relation of *compatibility*, $C \subseteq W \times W$, between worlds (see Dunn 1993), whereby the negation of A holds at world w just in case, at all compatible worlds, A fails to hold. The clause for negation is then:

(SC¬) $v_w(\neg A) = 1$ if for all $w_1 \in W$ such that Cww_1, $v_{w_1}(A) = 0$, and 0 otherwise.

On this definition, ¬ is a 'negative modality': a quantifier over worlds, restricted by an accessibility relation interpreted as compatibility. Because we utter negations to express incompatibilities and exclusions (Berto 2015), a semantics for negation grounded in compatibility makes intuitive sense.

Restall (1999) then shows how to get the Routley Star negation (S¬) from §6.1 out of (SC¬), by imposing conditions on compatibility. It must be symmetric (if Cww_1 then Cw_1w) and serial (for all $x \in W$, there is a $y \in W$ such that Cxy). And each world w must have a maximal compatible world: some $x \in W$ such that Cwx and, for all y, if Cwy then $y \preccurlyeq x$. Restall then claims that

> given that the compatibility semantics makes sense and is an applied semantics, it follows that its simple retelling, involving the Routley star, also makes sense, and it too is an applied semantics. (Restall 1999, 63)

These various ways of thinking about R in terms of information have been popular, and they allow for interplay between relevant logic

and other theoretical frameworks for reasoning about information, such as situation semantics. But let us briefly mention a worry for such interpretations. (We now draw on Jago 2013d.)

Information, as the term is used by Mares at least, has to be cognitively accessible, for 'what counts as a situation depends on the discriminatory capacities of human beings' (Mares 2009, 350). So, if w carries the information that A, then it should be possible for someone in w to get the information that A, and hence to come to know that A. But this idea is in tension with Excluded Middle, $A \vee \neg A$, which is valid in **R**. (It's also valid in other strong relevant logics, including Anderson and Belnap's favourite system **E**, which adds necessity to the relevant conditional.) But there's no reason to think that there is a situation in which, for every A, either the information that A or the information that $\neg A$ is available.

One can reply, correctly, that the semantics for **R** or **E** doesn't require *every* situation to support Excluded Middle. In fact, as we have seen in §6.1, to avoid irrelevancies it is vital that some points be inconsistent (i.e., both A and $\neg A$ hold there, for some A), and some be incomplete (i.e., neither A nor $\neg A$ holds there). But the objection is that (for logics like **R** and **E**) the semantics requires all *normal* points to support Excluded Middle. And yet, on the current interpretation, it is unlikely that there exist such situations. (The objection does not threaten weaker relevant logics in which Excluded Middle is not a theorem.)

6.4 Conditionality Interpretations

Beall et al. (2012) take a different approach to interpretations of R in the ternary semantics. (We follow the presentation in Jago 2013d in this section.) Beall et al. argue that, whichever way we think of conditionality in general, we get a suitable interpretation of R. They consider three general ways of thinking about conditionality:

(6.13) as the exclusion of counterexamples;

(6.14) as an operator or function; and

(6.15) as the kind of notion supported by conditional logic.

We'll consider options (6.13) and (6.14) only here, for the relevant arrows considered in this chapter are not connected to the kind of *ceteris paribus* conditionals studied in conditional logic.

Suppose, along with reading (6.13), we think of conditionality as the ruling out of certain situations: $A \rightarrow B$ says that there are no A-situations which are not also B-situations. That's a very classical way of thinking about the conditional, where a counterexample to $A \rightarrow B$ is any situation where A is true but B is false. We can also think of a strict conditional $A \dashrightarrow B$ (§1.2), understood as the necessitation of a material conditional, $\Box(A \supset B)$, in these terms: w_1 is a counterexample to $A \dashrightarrow B$ at w when w_1 is accessible from w and A but not B is true at w_1.

The difficulty with running this interpretation in the case of the ternary relation semantics is that the points of evaluation of antecedent and consequent may differ. To check whether $A \rightarrow B$ holds at w, we need to check for A at w_1 and B at w_2 whenever Rww_1w_2. A counterexample to $A \rightarrow B$ at w, therefore, depends on what goes on at some pair of points w_1 and w_2. In just this way, Beall et al. (2012) propose to treat 'split points' $\langle\!\langle w_1, w_2 \rangle\!\rangle$ as potential counterexamples. Truth and falsity at a split point $\langle\!\langle w_1, w_2 \rangle\!\rangle$ are fixed by truth at w_1 and falsity at w_2, respectively. Rww_1w_2 then says that $\langle\!\langle w_1, w_2 \rangle\!\rangle$ is accessible from (or possible relative to) w. So, just as in the modal strict conditional case, $\langle\!\langle w_1, w_2 \rangle\!\rangle$ is a counterexample to $A \rightarrow B$ at w when $\langle\!\langle w_1, w_2 \rangle\!\rangle$ is accessible from w, A is true at $\langle\!\langle w_1, w_2 \rangle\!\rangle$ but B is false there. We then get the relevant clause (S→) from §6.1 above.

If this approach is to provide a philosophical interpretation of R, as opposed to a useful bit of pure semantics, then the notion of a split point must be well understood. Notice that $\langle\!\langle w_1, w_2 \rangle\!\rangle$ cannot in general be thought of as the pair set of points $\{w_1, w_2\}$, for this is identical to the pair set $\{w_2, w_1\}$. But not so for split points, in which the order of the points is essential, as the definitions of truth and

falsity at a split point make clear. For the same reason, we can't think of $\langle\!\langle w_1, w_2 \rangle\!\rangle$ as the mereological composition or fusion of w_1 and w_2, for the fusion of thing a and thing b is the same as the fusion of thing b and thing a. We might think of $\langle\!\langle w_1, w_2 \rangle\!\rangle$ as some sort of list or sequence of w_1 and then w_2. But what is a list or sequence of two situations, and in what sense are sentences true or false relative to such lists or sequences? In general, one needs a story on which w_1 and w_2, *taken in that order*, constitute a counterexample to $A \to B$, which does not assume that they constitute a counterexample when taken the other way around. This problem may not be insuperable, but more explanation is needed if this is to be a philosophically satisfying applied semantics.

Now let's turn to Beall et al.'s option (6.14) for interpreting R. This involves thinking of the conditional $A \to B$ in terms of an operator or function, taking us from A to B. Intuitionists might think of this in terms of a function from a proof of A to a proof of B, for example. More generally, suppose it makes sense to *apply* one situation w to another w_1, as if w were a function and w_1 an argument. And suppose the result of this application, $w(w_1)$, is in some sense contained in some other situation, w_2. Then we can set Rww_1w_2 whenever $w(w_1)$ is contained in w_2. This gives us an understanding of R in terms of functional application and containment relations between situations. It's easy to make sense of these notions when the points in the frame semantics are proofs, programs, sets of evidence, or other syntactic constructions. This provides good reason to think that intuitionists and other constructivists can make sense of the ternary relation in this way.

This way of interpreting R resurrects the worry we raised in §6.3 for information-theoretic interpretations, however. Suppose points w and w_1 are understood as the kinds of entities which can be applied functionally to one another, such as proofs or sets of evidence. What justification do we then have for thinking that there's some such w at which, for every A, either A or $\neg A$ holds? If there are no such points, then Excluded Middle cannot be valid and we will be unable to give semantics along these lines for strong relevant logics such as **R**.

6.5 The Truthmaking Interpretation

We follow Jago 2013d in this section. An important feature of the points w, w_1, w_2 in the ternary semantics is that they may be partial or incomplete: it may be that neither A nor $\neg A$ holds at some w. Such points have fairly natural interpretations in epistemic terms, as we have seen, that is, as information states, evidence, or proofs. But such interpretations lead to problems in justifying Excluded Middle (§6.3). This suggests that a non-epistemic interpretation of partial points would be preferable, at least when considering strong relevant logics.

Restall (1996) suggests one such reading: the points are *truthmakers* (facts, states of affairs, or whatever else does truthmaking work). Restall briefly describes a truthmaker semantics which gives the first-degree fragment (i.e., without embedded conditionals) of the logic **RM**, a semi-relevant logic. (Van Fraassen (1969) had already spotted that a facts-based approach can give semantics for **FDE**, which we met in Chapter 5.)

This approach seems to us to overcome the Excluded Middle worry we raised in §6.3 for the epistemic, information-theoretic interpretation of worlds. A truthmaker for A will typically not be a truthmaker for B or for $\neg B$, unless there is some close relationship between A and B. So many truthmakers satisfy the partiality requirement. Yet plausibly, there are 'big truthmakers', such as complete-maximal worlds, which do make everything of the form $A \vee \neg A$ true. These can serve as the normal (validity-determining) points in the semantics. (We mentioned how one can stipulate that all normal points be maximal in our discussion of the Classicality Condition (CC) in §6.1 above. We saw there that (CC) poses no threat to the relevant requirement of the Variable Sharing Property.)

This approach has not been investigated in much detail, but it promises a philosophical interpretation of relevant logic in terms of familiar truth-like notions. It also suggests that principles of relevant logic are pertinent to the metaphysical debate over truthmaking. (The debate between Jago (2009a) and Rodriguez-Pereyra (2009) over certain metaphysical principles of truthmaking can be reinterpreted

as a semantic debate about whether $\wedge$ and $\vee$ are idempotent, so that $A \wedge A$ and $A \vee A$ are both equivalent to A, for example.) As Restall says, the approach is of interest 'to all those who seek to understand contemporary work on relevant logic, and for those who wish to form a robust theory of truthmaking' (Restall 1996, 339).

Chapter Summary

Relevant logics aim to avoid the 'paradoxes' of the material and strict conditionals. Their most natural semantics – the *Routley-Meyer semantics* – is given in terms of impossible worlds (§6.1). By placing certain further conditions on those worlds, we can obtain stronger relevant logics, including **TW** and **R** (§6.2). One of the main philosophical issues surrounding the general approach concerns how to interpret the Routley-Meyer ternary relation R on worlds and the Routley Star $*$. The information-theoretic interpretation has proved popular but, we argue, it faces philosophical issues (§6.3).

An alternative interpretation takes its cue from ways of thinking about conditionality in general. We considered the three options suggested by Beall et al. (2012), but found issues with each of them. A final option is the truthmaker interpretation of relevant logics, suggested by Restall (1996), which is promising but underdeveloped.

7

The Logic of Imagination

7.1 Hyperintensional Imagination

'Imagining' is highly ambiguous, as we saw in §1.5. We use the word for such different mental activities as daydreaming, hallucinating, supposing, planning, make-believing. We will focus on mental states with a propositional content (imagining *that A*: that Obama is blond-haired, that Holmes walks through Victorian London). We will target a notion found in such widely discussed works on imagination as Chalmers (2002b) and Yablo (1993), and dubbed by the latter 'positive conceivability'.

Positively conceiving that *A* is understood as a mental operation different from merely supposing or assuming that *A*, as when we make an assumption in a mathematical proof and, in some sense, as more substantive (Balcerak Jackson 2016). We represent a situation – a configuration of objects and properties – of which *A* is a truthful description:

> Positive notions of conceivability require that one can form some sort of positive conception of a situation in which *A* is the case. One can place the varieties of positive conceivability under the broad rubric of *imagination*: to positively conceive of a situation is to imagine (in some sense) a specific configuration of objects and properties. … Overall, we can say that *A* is positively conceivable when one can imagine that *A*: that is, when one can imagine a situation that verifies *A*. (Chalmers 2002b, 150, notation modified)

Similarly, Yablo (1993) speaks of conceiving that *A* as imagining a *world* verifying *A*. (Yablo grants that we do not imagine the world in all detail.) This seems to be the notion typically at issue in debates on whether conceivability entails possibility (see e.g. Gendler 2000, Hill 1997, Kung 2010, Roca-Royes 2011, Stoljar 2007, and the essays in Gendler and Hawthorne 2002). As we also saw in §1.5, in these debates it is often not clear what kind of mental representation is involved in the relevant act of conceiving or imagining. It is not clear whether it involves linguistic mental representations, or pictorial mental imagery mimicking corresponding sensory modalities. (We'll return to the issue in §7.3.)

Conceivability in the sense of Chalmers and Yablo seems to be linked to mental simulation, a phenomenon studied in cognitive psychology. We simulate alternatives to reality in our mind, in order to explore what would and would not happen if they were realized. This helps us to cope with reality itself, by improving future performance and allowing us to make contingency plans. (The works in Markman et al. 2009 explore various empirical phenomena in this ballpark.) That some things would happen in the envisaged scenario, and some would not, seems to imply that such exercises have some kind of logic: some things follow in the imagined situation, and some do not (Byrne 2005).

Works on the logic of imagination typically resort to a possible worlds framework, modelling imagination as a restricted quantifier over possible worlds (Costa Leite 2010, Niiniluoto 1985). But imagination, *qua* intentional mental state, is hyperintensional. Lois Lane can imagine that Superman is in love with her without imagining that Clark Kent is in love with her, as she ignores their being identical; we can imagine proving that $107 + 215 = 322$ without imagining proving Fermat's Last Theorem; and we can imagine that water turns out not to be H_2O (§1.5). This makes the phenomenon difficult to model via standard possible worlds semantics.

Wansing (2017) uses *neighbourhood semantics*, which we met in §5.2, for his logic of imagination. This allows several logical closure properties to fail for it: one's imagining that *A*, and that if *A* then *B*,

for instance, does not entail one's imagining that B. However, it is still the case that, as a consequence of the adoption of neighbourhood semantics, if A is logically equivalent to B and one imagines that A, one automatically imagines that B and vice versa. But this result seems wrong. Even in weak logics such as the basic relevant logic **B** of §6.1, A is equivalent to $A \vee (A \wedge B)$. However, one may imagine that A without imagining that $A \vee (A \wedge B)$, for every B. One may lack the very concepts involved in B, for example.

Impossible worlds are thus natural candidates for modelling imagination as mental simulation. But imagination, so understood, seems to have further features with which an acceptable model must comply. We'll describe them in the rest of this section, and present a simple impossible worlds semantics in the following §7.2, drawing on Berto (2014, 2017).

One feature of imagination as mental simulation is that it can be voluntary in ways belief cannot. One can imagine that all of one's home town has been painted yellow but, having overwhelming evidence of the contrary, one cannot easily make oneself believe it. Conscious acts of imagination as mental simulation can have an arbitrary, explicit starting point (Langland-Hassan 2016, Wansing 2017). This may be determined by the agent (as in, 'now let's imagine what would happen if …'), or it may be helped by external inputs (think of going through a novel, taking the sentences you read as your explicit input). According to Nichols and Stich's influential *mental simulation model* (2003), we begin imagining with 'an initial premiss or set of premisses, which are the basic assumptions about what is to be pretended' (2003, 24).

Imagination is not purely inferential, however. 'Children and adults elaborate the pretend scenarios in ways that are not inferential', filling in the explicit instruction with 'an increasingly detailed description of what the world would be like if the initiating representation were true' (Nichols and Stich 2003, 26–8; see also Langland-Hassan 2016, van Leeuwen 2016). You read a Jeffery Deaver book featuring Lincoln Rhyme, a detective working in New York on some murder case. The sentences of the book give you the explicit input. You integrate it

with background information you've imported into the scenario, on the basis of what you know or believe: New York is in the US, and normally detectives are human beings, although (let's suppose) the Deaver story does not state these things explicitly. Absent information to the contrary, you imagine Lincoln as a human being working in the US, although this is not entailed logically by the explicit input.

We propose to model this via modal operators interpreted as *variably* strict quantifiers over worlds, possible or otherwise. The variability of the quantifiers accounts for the contextual selection of the information we import in acts of imagination. As we will see, the input will play a role similar to a conditional's antecedent in Lewis's (1973b) semantics for counterfactuals.

It's important, however, not to treat agents as importing too much background information into acts of imagination. We do not indiscriminately import arbitrary, unrelated contents into imagined scenarios. You know that Manila is the capital of the Philippines, but this is immaterial to your imagining Lincoln Rhyme's New York adventures. Such adventures do not involve Manila or the Philippines at all. So you will not, in general, import such irrelevant content in your scenario. Of course, you *can* imagine things about Manila as well, by some free-floating association of ideas; but you will avoid it while engaging in mental simulation specifically of Lincoln Rhyme's New York adventures. So such exercises of imagination must obey some constraint of relevance.

7.2 A Semantics of Imagination

We will use a propositional language $\mathscr{L}$ with the usual set of atoms AT closed under negation $\neg$, conjunction $\wedge$, disjunction $\vee$, a strict conditional $\neg\!\!>$, modal operators $\Box$ and $\Diamond$, and square brackets '[' and ']', put to special use. The well-formed formulas are the atoms and, if A and B are well-formed formulas, then so are:

$$\neg A \mid (A \wedge B) \mid (A \vee B) \mid (A \neg\!\!> B) \mid \Box B \mid \Diamond B \mid [A]B$$

Things of the form '[A]' are modal operators indexed by formulas. (In conditional logic, this idea goes back to Chellas (1975).) Take a bunch of acts of imagination, performed by a given agent on specific occasions, and characterized by an explicit input: what the agent sets out to imagine ('Let's imagine that Holmes chases Moriarty across London in a horse-drawn carriage'), which can be taken as corresponding to Nichols and Stich's 'initial premiss', the 'basic assumption about what is to be pretended'. This is given directly by a formula of $\mathscr{L}$. If K is the set of formulas expressing possible explicit inputs, then for each $A \in K$, [A] is the corresponding modal. (K might be the whole language, or the language free of the '[A]' operators, or some restricted fragment of it. Just which restrictions should be put on K is an interesting issue, which we will not pursue here.) We can read '[A]B' as 'it is imagined in the act with explicit input A, that B'; or, more tersely, 'it is imagined in act A that B'. We will call each [A] an *imagination operator*.

The semantics is inspired by the relational frames for **FDE** (§5.4). An *imaginative frame* $\mathscr{F}$ is a triple $\langle\!\langle W, N, \{R_A \mid A \in K\}\rangle\!\rangle$. W is the set of worlds; $N \subseteq W$ is the subset of normal worlds; the worlds in $W-N$ are the non-normal or impossible worlds; and each $R_A \subseteq W \times W$ is a binary accessibility relation on W, one for each sentence $A \in K$.

A frame becomes a model $\mathscr{M} = \langle\!\langle W, N, \{R_A | A \in K\}, \rho\rangle\!\rangle$ when endowed with a valuation relation ρ, relating (for each world w) the atoms in AT to truth ('$\rho_w p1$'), falsity, ('$\rho_w p0$'), both, or neither. We then extend ρ to the whole language as follows. For the extensional connectives we have, for all $w \in N$:

(S1¬) $\rho_w(\neg A)1$ iff $\rho_w A0$

(S2¬) $\rho_w(\neg A)0$ iff $\rho_w A1$

(S1∧) $\rho_w(A \wedge B)1$ iff $\rho_w A1$ and $\rho_w B1$

(S2∧) $\rho_w(A \wedge B)0$ iff $\rho_w A0$ or $\rho_w B0$

(S1∨) $\rho_w(A \vee B)1$ iff $\rho_w A1$ or $\rho_w B1$

(S2∨) $\rho_w(A \vee B)0$ iff $\rho_w B0$ and $\rho_w B0$

The familiar modalities get their usual **S5** clauses, over normal worlds. For all $w \in N$:

(S1$\neg\exists$) $\rho_w(A \neg\exists B)1$ iff for all $w_1 \in N$, if $\rho_{w_1}A1$, then $\rho_{w_1}B1$

(S2$\neg\exists$) $\rho_w(A \neg\exists B)0$ iff for some $w_1 \in N$, $\rho_{w_1}A1$, and $\rho_{w_1}B0$

(S1$\Box$) $\rho_w(\Box A)1$ iff for all $w_1 \in N$, $\rho_{w_1}A1$

(S2$\Box$) $\rho_w(\Box A)0$ iff for some $w_1 \in N$, $\rho_w A0$

(S1$\Diamond$) $\rho_w(\Diamond A)1$ iff for some $w_1 \in N$, $\rho_{w_1}A1$

(S2$\Diamond$) $\rho_w(\Diamond A)0$ iff for all $w_1 \in N$, $\rho_w A0$

As for the $[A]$s, for $w \in N$:

(S1$[A]$) $\rho_w([A]B)1$ iff for all $w_1 \in W$ such that $R_A w w_1$, $\rho_{w_1}B1$

(S2$[A]$) $\rho_w([A]B)0$ iff for some $w_1 \in W$ such that $R_A w w_1$, $\rho_{w_1}B0$

Read '$R_A w w_1$' as saying that w_1 is accessed by an act of imagination with explicit input A, performed at w.

These recursive truth conditions have been defined for worlds in N. For worlds in $W - N$, ρ relates all formulas directly to truth values, irrespective of their syntax. (We met this approach to non-normal worlds in Rantala frames for epistemic logic, §5.3.) Logical validity and consequence are defined, once again, as truth and truth preservation (respectively) at all normal worlds in all models.

At normal worlds, the recursive truth and falsity conditions for $[A]$-formulas can also be expressed using set-selection functions, as in Lewis's (1973b) semantics for counterfactuals. Each $A \in K$ comes with a projection function f_A, taking as input the world where the act is performed and giving the set of worlds made accessible: $f_A w = \{w_1 \in W \mid R_A w w_1\}$.

A key idea in Lewis's semantics is that f_A outputs the A-worlds that are most objectively similar to the input world. In the context of a semantics for imagination, we can take it as outputting the worlds

that are most subjectively plausible for the agent, given input A. Imagination, on this model, works as a kind of belief revision. We will not impose a relation of comparative plausibility for worlds, as Baltag and Smets (2006), Grove (1988), and Leitgeb and Segerberg (2007) do. How orderings of this kind should work when impossible worlds are around involves a number of open issues, and the semantics below is largely independent of them. We will come back to plausibility orderings involving impossible worlds in §11.3, in the context of a treatment of truth in fiction.

Let $|A|$ be the set of worlds where A is true. Then, for $w \in N$:

(S1[A]) $\rho_w([A]B)1$ iff $f_A w \subseteq |B|$

(S2[A]) $\rho_w([A]B)0$ iff $f_A w \cap |\neg B| \neq \varnothing$

So for normal worlds, $[A]B$ is true (false) at w iff B is true at all worlds (false at some world) in a set selected by f_A. The relational and functional clauses are equivalent, given that $R_A w w_1$ iff $w_1 \in f_A w$. But it will sometimes be easier to talk using one formulation rather than the other.

A natural constraint on the semantics is that, for all $A \in K$ and $w \in N$:

(OBTAINING) If $w \in N$, then $f_A w \subseteq |A|$

Normal worlds access only those worlds where the explicit imaginative input *obtains*. Thus, in an act of imagination with explicit input A, one looks only at worlds where A is true. (We restrict this to normal worlds, since the non-normal/impossible worlds can do what they like.) We will consider only models satisfying OBTAINING.

To represent the imaginability of inconsistencies, we have allowed formulas to be related both to truth and to falsity (or to neither). But we may not want this to happen at normal worlds. We can add to the semantics a Classicality Condition as we did for relevant logics in §6.1, requiring all normal worlds to be maximally consistent with respect to atoms:

(CC′) If $w \in N$, then for each $p \in AT$, either $\rho_w p1$ or $\rho_w p0$ but not both.

This generalizes to all formulas not including imagination operators (as an easy induction on the complexity of formulas shows). To extend it to the whole language, we would need a different falsity condition (S2[A]). One option is to take $[A]B$ to be false when it is not true: for $w \in N$,

(S2[A]′) $\rho_w([A]B)0$ iff it is not the case that $\rho_w([A]B)1$.

This prevents inconsistencies accessible via imagination in non-normal worlds from creating inconsistencies at normal worlds.

Once the framework has been 'classicalized' in this way, the logic induced by the semantics for the connectives other than the imagination operators is just normal propositional **S5**. Let us now move on to an exploration of what is and what isn't valid in the semantics.

7.3 The Mereology of Imagination

Our first logical validity is:

$$\vDash [A]A$$

The explicit input is always imagined. This is immediately guaranteed by OBTAINING. Next come validities that involve conjunction. Firstly, for reasons to be discussed soon, we may want that, when one imagines that a conjunction is the case, one imagines each conjunct. The following condition does the trick:

(SIMPLIFICATION) For all $w \in N$: if $R_A w w_1$, and $\rho_{w_1}(B \wedge C)1$, then $\rho_{w_1} B1$ and $\rho_{w_1} C1$

This gives the following validities:

$$[A](B \wedge C) \vDash [A]B \qquad [A](B \wedge C) \vDash [A]C$$

To see why these follow, suppose $w \in N$ and $\rho_w([A](B \wedge C))1$. By (S1[A]), for all R_A-accessible worlds w_1, $\rho_{w_1} B \wedge C1$. Then SIMPLIFICATION gives us $\rho_{w_1} B1$ and $\rho_{w_1} C1$ and so, by (S1[A]) again, we have $\rho_w[A]B1$ and $\rho_w[A]C1$.

The companion constraint to SIMPLIFICATION is:

(ADJUNCTION) For all $w \in N$: if $R_A w w_1$, $\rho_{w_1} B1$, and $\rho_{w_1} C1$, then $\rho_{w_1}(B \wedge C)1$

This gives us:

$$[A]B, [A]C \vDash [A](B \wedge C)$$

To see why, suppose $w \in N$, $\rho_w([A]B)1$ and $\rho_w([A]C)1$. By (S1[A]), for all R_A-accessible worlds w_1, we have $\rho_{w_1} B1$ and $C\rho_{w_1} C1$. By ADJUNCTION, $\rho_{w_1}(B \wedge C)1$ for any such w_1 and so, by (S1[A]), $\rho_w([A](B \wedge C))1$.

Imaginative accessibility is then *de facto* limited by SIMPLIFICATION and ADJUNCTION to worlds that behave with respect to conjunction. But ADJUNCTION may be found problematic. Is it so that, when one imagines that B and that C in a single act [A], one automatically imagines their conjunction?

A similar question has been asked for counterfactuals, namely whether different counterfactuals with the same antecedent demand the conjunction of their consequents, given the role consequents play in fixing the context of evaluation. Suppose one sets out to imagine Caesar, the Roman emperor, being in command of the US troops in the Korean War. (The example is Quine's (1960, 222).) This gives an explicit input, A. One can then unfold the scenario as one where Caesar uses the atom bomb, $[A]B$, if one imports into the representation information concerning the weapons available in the twentieth century. Or, one can get to imagine him using catapults, $[A]C$, if one rather allows the Roman military apparatus to step in. However, we shouldn't thereby infer $[A](B \wedge C)$, involving Caesar's employing both bombs and catapults. One *can* imagine that, too, if one likes, but it should not come as an automatic logical entailment.

We think that something has gone wrong with the reconstruction of the situation. Acts of imagination are contextually determined, in that the same explicit input can trigger the importation of different background information in different contexts (the time and place where the act takes place, the status of the agent's background information, and so on). In Quine's example, there is a clear contextual shift. So ADJUNCTION can be maintained by adding a contextual parameter. This can be represented in the formalism by a set of contexts and variables ranging over them, indexing imaginative acts. $[A]_x$ and $[A]_y$ will then stand for two distinct acts with the same explicit input A. Once the adjunctive inference is parameterized to contents with the same index, it avoids the worry in Quine's example.

We think ADJUNCTION is important in modelling an independently plausible conception of imagination. Imagination, in the relevant sense, is more than mere supposition of a content, A. Rather, it is positive conceivability in Chalmers's (2002b) sense: someone's representing a scenario or a state of affairs which makes A true. It is a generally agreed principle of the logic of (inexact) truthmaking (Fine 2014, Yablo 2014) that truthmakers behave in such conjunctive fashion. This mirrors the idea that states of affairs themselves can stand in parthood relations. (A conjunctive state of affairs that object m is P and object o is Q includes as constituents the individual states of affairs that m is P and that o is Q. We discuss logically complex states of affairs in Barker and Jago 2012 and Jago 2011.)

If imagination crucially involves mental imagery (as Kind (2001) argues), then there may be a certain mereological structure implemented in the mind, and we may find evidence for this in empirical psychology. (We mentioned the issue back in §1.5.) Empirical evidence for the quasi-spatial features of pictorial mental imagery has been gathered since the 1970s, including experimental work on mentally scanning images (Block 1983, Pinker 1980, Shephard and Metzler 1971). Such work showed, for instance, that the time taken to scan between two points of a mental image is often proportional to their subjective distance in the pictorial representation; that larger objects 'fill' the space sooner than smaller ones; and that the level of detail of the

depicted situations decreases at the periphery of the image, similarly to what happens with our visual field.

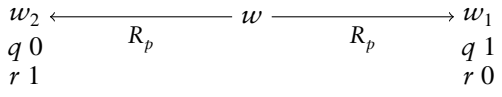
The view is disputed, however. Zenon Pylyshyn (1973, 1981, 2002) accepts the existence of mental images, but claims that their representational features can be reduced to non-pictorial, linguistically encoded representations. This 'imagery debate' is one of the most intractable controversies in contemporary cognitive psychology. If it resolves in favour of the quasi-spatial features of pictorial mental imagery, then we should include both SIMPLIFICATION and ADJUNCTION in the semantics. If Pylyshyn is right, and mental imagery can be reduced to linguistic representation, then we may want to drop SIMPLIFICATION or ADJUNCTION (or both).

7.4 The Under-Determinacy of Imagination

We move on to issues concerning disjunction. Imagination generally under-determines its contents: we imagine things vaguely, without this entailing that we imagine vague things. When you imagine Sherlock Holmes, by default you imagine him either left-handed or right-handed (or ambidextrous); but it's perfectly possible to not imagine him left-handed and not imagine him right handed either. Our semantics captures this:

$$[A](B \vee C) \not\models [A]B \vee [A]C$$

For a countermodel, take three normal worlds w, w_1, w_2:



Every world R_p-accessible from w verifies $q \vee r$ and so w verifies $[p](q \vee r)$. But since w_1 doesn't verify r and and w_2 doesn't verify q, w verifies neither $[p]q$ nor $[p]r$.

Notice that this countermodel does not require non-prime worlds, which make a disjunction true without making either disjunct true. All

worlds in the countermodel are normal. (This isn't required, however: only w must be a normal world for the countermodel) The underdeterminacy of imagination is delivered by the plurality of worlds accessed via R_A. Each accessible world may fill in the unspecified details in different ways. Since each $[A]$ is a universal modal operator, differences between the accessible worlds translate into unspecific acts of imagination.

We need worlds where disjunction behaves non-normally for another reason, however. When one imagines that B, one need not thereby imagine the disjunction of B and any arbitrary C (§7.1). One may lack the concepts involved in C. Our semantics captures this feature, too:

$$[A]B \not\models [A](B \vee C)$$

For a countermodel, we need a normal world w and non-normal world w_1:

$$w \xrightarrow[R_p]{} w_1$$
$$q\ 1$$
$$\not{q \vee r}\ 1$$

(Here, we've written $\not{q \vee r}\ 1$ to show that w_1 doesn't verify $q \vee r$. This is possible because w_1 is a non-normal world: ρ_{w_1} can relate any formula to any value or none.) In the model, by (S1[A]), w verifies $[p]q$ but not $[p](q \vee r)$. If we use our original clause (S2[A]) for falsity, then $[p]q \vee r$ is neither true nor false at w. But if we use the revised clause (S2[A]′), then $[p]q \vee r$ is false at w. Either way, the countermodel requires w_1 to be non-normal, else q's being true would force $q \vee r$ to be true.

7.5 Non-Monotonicity and Relevance

Imagination operators are non-monotonic, in the following sense:

$$[A]B \not\models [A \wedge C]B$$

To see why, consider this counterexample, with w a normal world:

$$w \xrightarrow[R_{p \wedge r}]{} w_1$$

Since w_1 does not verify q, w does not verify $[p \wedge r]q$. But since no world is R_p-accessible, w trivially verifies $[p]A$ for every A, including $[p]q$. So $[p]q \not\models [p \wedge r]q$.

As an act of imagination (in a given context) is individuated by its explicit content, one cannot in general import information into the explicit content itself without turning it into a different act. As you imagine Holmes walking across London, you imagine him walking across a British city. But if you imagine Holmes walking across London and that London has been displaced to France, you will not imagine him walking across a British city.

Other invalidities highlight the hyperintensional nature of imagination:

$$A \mathrel{\neg\!\!\exists} B \not\models [A]B$$

Here's the counterexample, with w a normal and w_1 a non-normal world:

$$\begin{array}{l} w \\ q\ 1 \end{array} \xrightarrow[R_p]{} w_1$$

At every normal world (just w), if p then q, and hence w verifies $p \mathrel{\neg\!\!\exists} q$. (Recall that the strict conditional $\mathrel{\neg\!\!\exists}$ looks at normal worlds only.) But this does not guarantee that, in imagining that p, we must thereby imagine that q. What we imagine may be impossible, as it is at w_1.

In particular, strict irrelevant conditionals fail the Variable Sharing Property from §6.1. But those conditionals do not entail the corresponding irrelevant imaginings. Here's an irrelevant but valid strict conditional:

$$\models (A \wedge \neg A) \mathrel{\neg\!\!\exists} B$$

But the corresponding imagination formula isn't valid:

$$\not\models [A \wedge \neg A]B$$

and so

$$(A \wedge \neg A) \mathrel{-\!\!\!\circ} B \not\models [A \wedge \neg A]B$$

Here's the counterexample, with w a normal and w_1 a non-normal world:

$$w \xrightarrow[R_{(p \wedge \neg p)}]{} w_1 \qquad q\ 1$$

Imagining an inconsistent scenario does not trivialize our act of imagination. We can discriminate between different logical necessities and we do not imagine them all automatically whenever we explicitly imagine something. Thanks to non-normal worlds we have:

$$\Box B \not\models [A]B \qquad\qquad \neg\Diamond A \not\models [A]B$$

A constraint that should *not* hold in our semantics is (the counterpart of) what Lewis (1973b) called 'Weak Centring':

(WEAK CENTRING) If $w \in |A|$, then $w \in f_A w$.

This entails that, if a world w realizes the explicit content of an act of imagination A, then w is one of the worlds in the set outputted by the selection function for A. Even restricted to normal worlds, WEAK CENTRING validates a sort of modus ponens for imagination:

$$A, [A]B \models_? B$$

According to this principle, if the explicit content A of an act of imagination is true, and it is imagined in that act that B, then B also true. But this is wrong. Franz and Mark imagine that, in setting themselves a writing deadline, frenetic and productive writing will ensue ($[A]B$). They set the deadline (A); but things go on much as before for them ($\neg B$). Well-intentioned Brenda imagines, in her

country leaving the EU, things getting better for her ($[A]B$). Her county does leave (A), but things get worse for Brenda ($\neg B$). So we should not accept WEAK CENTRING.

One way WEAK CENTRING can go wrong is when an agent imports false beliefs into her imaginings. We have spoken of importing *information* into imaginings, but we shouldn't assume that what's imported must be true. (If you think information is by definition factive, then we're speaking of non-factive quasi-information.) 'What people do not change when they create a counterfactual alternative [in imagination] depends on their beliefs' (Byrne 2005, 10), and false beliefs may sneak in. You imagine Merkel signing treaties in Brussels ($[A]$), but you mistakenly believe Brussels to be in France. You import that belief and imagine Merkel to be signing treaties in France ($[A]B$). Merkel does in fact sign in Brussels (A), but this doesn't imply Brussels is in France. In general, the role of the world w where the act of imagination takes place is to fix the agent's beliefs, rather than to fix what is in fact the case.

7.6 Imaginative Equivalents

So far our logic of imagination is relatively weak, due to its highly hyperintensional features and the variability in the selection of the accessible worlds. We may add a *Principle of Imaginative Equivalents* (mimicking an analogous principle that holds in conditional logics (Priest 2008, 92)), whose effect is to limit the hyperintensionality of imagination:

(PIE) If $f_A \subseteq |B|$ and $f_B \subseteq |A|$, then $f_A w = f_B w$.

This says: if all the selected A-worlds make B true and vice versa, then A and B are 'imaginative equivalents'. When we set out to imagine either, we look at the same set of worlds. Given (PIE), we have:

$$[A]B, [B]A, [A]C \vDash [B]C$$

To see why, suppose normal world w verifies $[A]B, [B]A, [A]C$. By (S1[A]), we have $f_A w \subseteq |B|$, $f_B w \subseteq |A|$, and $f_A w \subseteq |C|$. (PIE) then gives us $f_A w = f_B w$ and hence $f_B w \subseteq |C|$. So by (S1[A]), w verifies $[B]C$.

This inference tells us that 'imaginative equivalents' A and B can be replaced *salva veritate* as modal indexes in $[\cdot]$. Given the number of hyperintensional distinctions we can make in our imagination, there may be few imaginative equivalents for a given agent. But suppose that *bachelor* and *unmarried man* are imaginative equivalents for you (as they should be, we guess, for any competent English speaker): you are so firmly aware of their meaning the same that you cannot imagine someone being and not the other. So when you imagine that John is unmarried, you imagine that John is a bachelor ($[A]B$) and vice versa ($[B]A$). Suppose, in imagining that John is a bachelor, you imagine that he has no marriage allowance ($[B]C$). Then the same happens when you imagine that he is unmarried, $[A]C$.

(PIE) also licenses an inference we'll call *Special Transitivity*:

(ST) $[A]B, [A \wedge B]C \vDash [A]C$

To see why, suppose normal world w verifies $[A]B$ and $[A \wedge B]C$. Given OBTAINING, w also verifies $[A]A$ and so, by ADJUNCTION, $[A](A \wedge B)$. From OBTAINING and SIMPLIFICATION, we have $[A \wedge B](A \wedge B)$ and hence $[A \wedge B]A$ at all worlds. Then $f_A w \subseteq |A \wedge B|$ and $f_{A \wedge B} w \subseteq |A|$ and so, by (PIE), $f_A w = f_{A \wedge B} w$. Since w also verifies $[A \wedge B]C$, we have $f_{A \wedge B} \subseteq |C|$, hence $f_A w \subseteq |C|$, and so w verifies $[A]C$ too.

Should we accept (ST)? Some instances seem good. You imagine, in winning the lottery, having a lot of money ($[A]B$). You then imagine, in winning the lottery and having lots of money, that you'll have to pay a substantial amount of tax ($[A \wedge B]C$). It seems OK to infer *in the same context* that, in imagining winning the lottery, you thereby imagine having to pay a substantial amount of tax ($[A]C$). Of course, you can also imagine winning the lottery and avoiding paying any taxes. But that seems to create a different context from the one in which you imagined having to pay the taxes, on winning the lottery and so having lots of money.

It may be that there are intuitive counterexamples to (ST) forceful enough for us to reject it. If so, we must drop either (PIE) or one (or both) of SIMPLIFICATION and ADJUNCTION. Perhaps what we should take from a counterexample to (ST), should one exist, is that imagination does not have a nature which respects conjunction after all (§7.3). If that's right, then we should drop either SIMPLIFICATION or ADJUNCTION (or both), without which the proof of (ST) will not go through, and retain (PIE) for the additional inferential power it gives us.

Chapter Summary

Imagination seems to have a logic, albeit one which is hyperintensional and sensitive to context (§7.1). We offered a semantics of imagination, with operators expressing 'imaginative acts' of mental simulation (§7.2). We then discussed a number of conditions we could impose on the semantics, in order to validate certain inferences (§7.3). One important issue is how acts of imagination interact with disjunction. One can imagine some disjunction as obtaining without being imaginatively specific about which disjunction obtains, for example (§7.4). We then turned to the issue of non-monotonicity: how B may follow from imagining that A, but not from imagining that $A \wedge C$ (§7.5). Finally, we discussed the *Principle of Imaginative Equivalents*, which, if valid, adds considerable power to the logic (§7.6).

Part III

Philosophical Applications

8

Hyperintensionality

8.1 Is Hyperintensionality Real?

In §1.3, we introduced the concept of *hyperintensionality*. We said that an operator $\mathscr{H}$ is hyperintensional when $\mathscr{H}A$ and $\mathscr{H}B$ can differ in truth value, even when A and B are necessarily (logically, mathematically, or metaphysically) equivalent. We identified concepts which have been given an intensional possible worlds account, such as knowledge, belief, meaning and semantic content, propositions, information, counterfactual conditionals. We claimed that these concepts are hyperintensional and hence that their possible worlds account falls short of capturing the concept.

Some philosophers deny that some of these concepts are genuinely hyperintensional. Some deny that there are any genuinely hyperintensional concepts. These denials sometimes derive from other theoretical commitments. A philosopher who analyses content or meaning in terms of possible worlds will be unable to find any content or meaning in hyperintensional operators. She will view any purported hyperintensional operators as meaningless. But since knowledge, belief, and so on are meaningful concepts, she will deny that these are hyperintensional. She'll then argue that the appearance of hyperintensionality (in examples involving knowledge and belief reports, for example) are illusory. She may also offer a positive analysis of the concept, showing why it can't be hyperintensional.

In the cases of knowledge and belief, we find these moves in Lewis (1982, 1986b) and Stalnaker (1984). Both authors support the

possible worlds analysis of knowledge, belief, and content. They base their support on considerations of the nature of these concepts. Both acknowledge that we do not *seem* to know all consequences of what we know, that logical reasoning *seems* to be informative, and so on. But, they argue, the appearances are misleading. We'll discuss their suggestions in §8.2.

Belief and knowledge are everyday concepts, and so we may draw on everyday reflections on those concepts as (defeasible) evidence for hyperintensionality. Proposition, essence, grounding, and (to some extent) meaning and content are, by contrast, terms of art. A philosopher has more scope to rebut purported examples of hyperintensionality using these concepts. 'That's not how I use that concept', she may say. She may even reject the concept outright. So our claims here will be somewhat conditional in nature. If one wants to adopt one of these concepts (in the way that we think is the most philosophically useful), then one had better accept hyperintensionality. We'll discuss these cases in §8.3.

8.2 The Epistemic Case

For Stalnaker, the nature of belief, as a rational attitude, rules out a hyperintensional analysis. He begins with an 'impressionistic' picture of human representational states:

> Representational mental states should be understood primarily in terms of the role that they play in the characterization and explanation of action. What is essential to rational action is that the agent be confronted, or conceive of himself as confronted, with a range of alternative possible outcomes of some alternative possible actions. The agent has attitudes, pro and con, toward the different possible outcomes, and beliefs about the contribution which the alternative actions would make to determining the outcome. One explains why an agent tends to act in the way he does in terms of such beliefs and attitudes. And, according to this picture, our conceptions of belief are conceptions of states which explain why a rational agent does what he does. (Stalnaker 1984, 4)

The key idea here is that human agents take *attitudes, pro and con, toward the different possible outcomes*. Suppose we assign a 1 to each possible outcome to which the agent takes a pro attitude, and a 0 when she takes a con attitude. Then the content of the attitude is captured as a *characteristic function* from possible outcomes to 1 or 0. Such functions are mathematically equivalent to the set of outcomes assigned 1. The content of any rational attitude is (equivalent to) a set of possibilities, on this analysis. Importantly, this argument doesn't proceed from a prior commitment to possible worlds semantics. Rather, it proceeds from an analysis of 'the role of beliefs and desires in the explanation of action' and concludes that 'the contents of those attitudes [must] distinguish between the alternative possibilities' (Stalnaker 1984, 23). Hence the possible worlds analysis.

Stalnaker acknowledges that we do not seem to be logically omniscient. The situation, he thinks, is that 'we have an argument to show that the identity conditions [on contents] are right, as well as examples that seem to show that they are wrong', and so 'the proper response is not so clear' (Stalnaker 1984, 24).

One suggestion he offers is a *metalinguistic* approach, on which

> the apparent failure to see that a proposition is necessarily true, or that propositions are necessarily equivalent, is to be explained as the failure to see what propositions are expressed by the expressions in question. (Stalnaker 1984, 84)

How might this explanation go? Stalnaker elaborates:

> Relative to any propositional expression one can determine two propositions: there is the proposition that is expressed, according to the standard rules, and there is the proposition that relates the expression to what it expresses. If sentence *S* expresses (according to the standard rules) proposition *p*, then the second proposition in question is the proposition that *S* expresses *p*. In cases of ignorance of necessity and equivalence, I am suggesting, it is the second proposition that is the object of doubt and investigation. (Stalnaker 1984, 84–5)

It seems that if I do not know a necessary truth, *A*, then what I do not know is that the sentence '*A*' expresses the proposition *that A*. I do

not know that '*A*' expresses the set of all possible worlds, according to Stalnaker. I countenance possibilities in which the words in '*A*' mean something other than their actual meaning and so, at some level, I fail to understand what '*A*' means. For

> Whenever the structure of sentences is complicated, there will be a nontrivial question about the relation between sentences and the propositions they express, and so there will be room for reasonable doubt about what proposition is expressed by a given sentence. (Stalnaker 1984, 84)

This is clearly a genuine phenomenon. Anyone who's read enough German, or even old-fashioned English, knows the feeling. You know what each word in turn means but, as the sentence runs on and on, you lose the overall meaning. You don't know what proposition is expressed. It happens. But we question whether it happens enough to make Stalnaker's case. Is it the case that, whenever one seemingly fails to know (or believe) a necessary truth, one is confused about what the sentence in question means? That seems implausible.

Here are two examples (from Jago (2014a)) to make the case. The first is mathematical. You learn, in school, about integers, equations, addition, and exponentiation. You don't learn all there is to know about them, of course, but enough to be comfortable with what those terms mean. (Let's say, you know how to relate exponentiation to multiplication to addition, and you know well what addition is.) You also know how to understand notation involving variables (quantified, implicitly, over integers). So you have no trouble understanding an equation like

$$x^n + y^n = z^n$$

Do you thereby know that, when $n > 2$, there are no three integers x, y, z fitting the equation? Of course not! It took mathematicians centuries to find a proof for that claim. Nevertheless, it's a mathematical and logical necessity (with 'logic' understood so as to contain basic arithmetic). In this case, Stalnaker must claim that you don't understand something about that statement. But, given the way we set up the example, that's hardly plausible.

The second example is more practical. Suppose you're playing chess (with no time controls) and agree to count a draw as a win for black. (Or suppose you're in a competition, and black needs only a draw to win the competition overall.) Here's a surprising mathematical fact: at each stage of the game, either you or your opponent has a *winning strategy* available. That's a function for generating the next move which, if followed to the letter, will result in victory (including a draw for black), regardless of what the other player does. A winning strategy exists right from the first move, either for white or for black (but we don't know which). If players always followed a winning strategy, then the same colour would always win (always white, or always black). In reality, that doesn't happen. Either the players don't know the winning strategy when it's there (they don't know that, say, Qe7 is the first move of a winning strategy), or else they're really not that fussed about a guaranteed win. (And it's obviously not the latter.)

In this case, Stalnaker again has to claim that we fail to know some of the meanings involved. But all that's required (mathematically speaking) to entail a winning strategy are (1) a precise specification of the rules of chess, (2) a precise specification of the conditions for winning, and (3) a precise description of the current state of the board. Given those facts, a description of a winning strategy (for one or other of the players) follows mathematically. So, the claim has to be that no chess player (who ever lost without wanting to) understands the meaning of the rules, the winning conditions, or the way we describe the state of the board. That is highly implausible.

In short, the phenomenon Stalnaker describes, not knowing which proposition is expressed by a given sentence, is genuine. But it is not able to account for many of the cases in which an agent seems not to know some necessary truth.

A different approach is to think of an agent's overall epistemic state as being split into *fragments*, or multiple 'frames of mind'. An agent may believe something in one frame of mind, something else in some other frame of mind, and never combine the two bits of information. She fails to believe all consequences of what she believes, on this

approach, because (as it were) she never puts two and two together. Lewis (1982) describes how to make sense of the phenomenon:

> I used to think that Nassau Street ran roughly east-west; that the railroad nearby ran roughly north-south; and that the two were roughly parallel. … So each sentence in an inconsistent triple was true according to my beliefs, but not everything was true according to my beliefs. … My system of beliefs was broken into (overlapping) fragments. Different fragments came into action in different situations, and the whole system of beliefs never manifested itself all at once. … The inconsistent conjunction of all three did not belong to, was in no way implied by, and was not true according to, any one fragment. That is why it was not true according to my system of beliefs taken as a whole. (Lewis 1982, 436)

On Lewis's approach, each fragment of belief can be treated in the possible worlds tradition. So, each fragment corresponds to a proposition, representing a consistent and logically closed set of beliefs. Yet the agent's beliefs in total are not closed under entailment and need not be consistent, as in Lewis's example. Stalnaker (1984) supports a similar view (which he holds in combination with the metalinguistic approach described above). Fagin and Halpern (1988) develop the idea into a formal semantics, which they call a model of 'local reasoning'. (The idea is that the logically omniscient 'reasoning' is local to each fragment, rather than a global feature of the agent's total belief state.)

Formally, the approach is similar to a non-adjunctive *subvaluational* semantics, on which premises A, B do not entail their conjunction, $A \wedge B$ (§4.4). On this approach, we consider a range of classical valuations, and consider a sentence to be true (simpliciter) when it's true-on-some-valuation. When A is true on one valuation and B true on another, but no valuation makes both A and B true simultaneously, we have a situation in which both A and B are true (simpliciter) but $A \wedge B$ is not. In the doxastic setting, this translates to the situation Lewis describes: he did not believe the inconsistent conjunction of the information he held about Nassau Street and the railroad.

As with the metalinguistic approach, the phenomenon described by this approach is genuine, but it cannot explain away enough cases of non-omniscience. Consider our chess example from above. Let's assume that our players are highly competent players, who hold all the relevant information (the rules, what it takes to win, and the current state of the game) in their current frame of mind. (Not all players are like that, of course. But many who reach a high level of competence are.) According to the fragments of mind approach, those players will know (and be able to act upon) a winning strategy. But, as we know from experience, that just isn't the case, not even for the very best players.

Following Jago (2014a), we think that the fragments of mind approach misdiagnoses why real agents aren't logically omniscient. According to the view, an agent doesn't believe a consequence of what she believes because she hasn't put the relevant premises together. But when she does put those premises together, she thereby, all of a sudden, comes to believe all their consequences. It's as if all the agent's deductive effort goes into combining the premises into a single belief state: from $A_1, \ldots, A_n$ to $A_1 \wedge \cdots \wedge A_n$. But that's the easy bit, deductively speaking. The hard bit, for real agents, is deriving further non-trivial consequences. Whether she derives these from the individual premises taken together, or from their conjunction, doesn't make that much difference. The information conveyed by a non-trivial valid deduction does not correspond to the move from individual premises to their conjunction, but rather in the deductive move from those premises (or their conjunction) to the conclusion. The fragments of belief approach can't account for this phenomenon.

There are also technical problems with the approach. Although an agent's total belief state is not closed under under entailment, on the fragments view, it is closed under single-premise entailment. If A entails B and she believes that A, then she automatically believes that B (and similarly for knowledge). This is condition (C1) from §5.1. This holds because each fragment is logically closed. The approach also validates (C2)–(C5), (C7), and (C8) from §5.1. These are all forms of logical omniscience we would like to avoid, both for belief

and for knowledge (with the exception of (C3) for knowledge, since one can't know inconsistent things). So, even on technical grounds, the fragments of belief approach isn't a sufficient solution to the problems of logical omniscience.

We've phrased our discussion in terms of logical omniscience. How does this relate to hyperintensionality? Non-hyperintensionality is one form of logical omniscience: (C4) from §5.1, *Closure under logical equivalence*. It says that knowledge (or belief) never distinguishes between logically equivalent contents. This is the very definition of a non-hyperintensional concept.

Closure under logical equivalence (C4) is closely related to closure under single-premise entailment. Suppose that A logically entails B. Then A is equivalent to $A \wedge B$. Given (C4), an agent who knows that A thereby knows $A \wedge B$ too. And if she knows the conjuncts of each conjunction she knows, then she knows that B too. (Similarly for belief.) The latter principle is *Closure under conjunction*, (C7). So (C4) plus closure under conjunction, (C7), implies (C1). If agents are not logically omniscient at all, then (C1) is false. But as we argued above, we cannot pin this failure on (C7), and so (C4) must be false too. So if agents are not logically omniscient, as we've been arguing, then (C4) is false, then knowledge and belief are hyperintensional concepts.

Here's the situation, as we see it. Examples such as our chess case intuitively show that agents aren't logically omniscient. Lewis's and Stalnaker's attempt to convince us otherwise fail. Absent a successful attempt along those lines, the evidence directs us to accept that knowledge and belief are genuinely hyperintensional.

8.3 The Case from Content

We've been discussing doxastic and epistemic contents. But the general notion of content is much broader. Every meaningful expression has a semantic content, not just those that are known, believed, or uttered. If epistemic or doxastic contents are hyperintensional, then *content*

in general is a hyperintensional notion. But there are independent (non-epistemic) reasons for thinking that this is the right way to think about content. The content of a truth-apt sentence is a proposition. We'll talk in terms of propositions, since there's a straightforward notion of logical equivalence for propositions. (We can also talk about logically equivalent subsentential terms, such as descriptions, but this would needlessly complicate our discussion.)

Both Lewis (1973b, 1986b) and Stalnaker (1976a, 1976b, 1984) take propositions to be sets of possible worlds. There's a very simple argument for that view. It's common to identify propositions with truth-conditions. But what, ontologically speaking, is a *condition*? Let's say we're interested in whether something meets condition X in such-and-such situations. We treat that condition as a function from the situations to the answers, *yes* or *no*. Mathematically, such functions are *characteristic functions*, and each such function defines a set, containing all and only the input entities for which the function's output is *yes*. It is then both very natural and mathematically elegant to identify the condition itself with the set of 'yes' situations. In the case of a truth-condition, the input situations are possible worlds and the outputs are *true* or *false*. So, on this very natural view, a truth-condition is a set of possible worlds. (This is similar to the argument we met at the start of §8.2.)

On this approach, the proposition $\langle A \rangle$ is the set of all possible worlds according to which it is the case that A. (Alternatively, the worlds which are such that, were they actual, then it would be the case that A.) This story comes with a ready-made account of what logically complex propositions are, for the set-theoretic account gives a Boolean algebra for propositions: $\langle A \wedge B \rangle$ is simply the set of worlds $\langle A \rangle \cap \langle B \rangle$ and similarly, $\langle A \vee B \rangle = \langle A \rangle \cup \langle B \rangle$, $\langle \neg A \rangle = \langle A \rangle^{c}$ (the set-theoretic complement of $\langle A \rangle$, on the domain of all possible worlds) and $\langle A \rightarrow B \rangle = \langle A \rangle^{c} \cup \langle B \rangle$. A consequence of this account is that necessarily equivalent propositions are identical (Stalnaker 1976a, 9). That is to say, the approach is intensional, not hyperintensional.

Despite the simplicity and naturalness of the approach, its shortcomings are many. We'll discuss two (related) ways to bring these out

(following Jago (2018b)). First, the *truthmaker objection*. Consider these two propositions:

(8.1) ⟨Lenny exists⟩

(8.2) ⟨Lenny exists ∧ 3 exists⟩.

True propositions have truthmakers. (Or, at the very least, true existential propositions have truthmakers.) Lenny alone truthmakes (8.1). But Lenny on his own does not truthmake (8.2): the number 3 gets in on the act, too. It's in virtue of the existence of both Lenny and the number 3, taken together, that (8.2) is true. Lenny is a *full* truthmaker for (8.1), but not for (8.2). Hence (8.1) stands in a relation to Lenny which (8.2) does not and so, by Leibniz's Law, (8.1) and (8.2) cannot be one and the same proposition. But they are necessarily equivalent and hence, according to the possible worlds view, one and the same proposition. Contradiction.

A related objection to the possible worlds account is the *aboutness objection*. *Aboutness* is 'the relation that meaningful items bear to whatever it is that they are *on* or *of* or that they *address* or *concern*' (Yablo 2014, 1). Research on aboutness has been flourishing recently, mainly thanks to Fine (2016) and Yablo (2014). Even before introducing a 'grand-sounding name for something basically familiar' (Yablo 2014, 1), logicians and semanticists were looking for a content-preserving entailment relations: relations that holds between two meaningful items *A* and *B* only when *B* introduces no content alien to what *A* is about. *Tautological entailment* (Van Fraassen 1969), *analytic containment* (Angell 1977, Correia 2004), and *analytic implication* (Parry 1933) are all variations on this theme.

As an example of how aboutness creates problems for the standard possible worlds account, consider the propositions:

(8.3) ⟨Lenny is stretching ∨ ¬Lenny is stretching⟩

which is about Lenny, not Bertie, and

(8.4) ⟨Bertie is barking ∨ ¬Bertie is barking⟩

which is about Bertie, not Lenny. They're about different things, and hence are different propositions. *Aboutness* should thus relate (8.3) to Lenny, not Bertie, and (8.4) to Bertie, not Lenny. So (8.3) and (8.4) stand in different relations, and hence are distinct entities. But they are logically equivalent. The possible world view then says they are one and the same proposition. Contradiction. In each case, we say, it's the possible worlds view of propositions that's at fault. It cannot model aboutness and aboutness-preserving entailment. There is broad agreement that a semantics that works properly for these notions should be hyperintensional (Yablo 2014, 62). (That said, Yablo himself works within the possible worlds framework, but with additional hyperintensional resources.)

We've argued that at least some concepts are hyperintensional. And unsurprisingly, we find the best way to account for such concepts is in terms of impossible (as well as possible) worlds. Impossible worlds give us a very natural and flexible way to account for hyperintensional concepts.

Let's continue with our focus on propositions. As we said above, it's common to identify propositions with truth conditions, and in general to identify conditions with characteristic functions. So a proposition is identified with the set of situations in which it is true. As we emphasized above, that's a very natural view. But there's nothing in this argument that says the situations in question have to be possible worlds. We take them to be sets of worlds, both possible and impossible. So we can accept the motivation whilst accepting that propositions are hyperintensional.

Just what logical properties such propositions have depends on the properties of impossible worlds. It depends on which logical rules, if any, impossible worlds must preserve. This is the *granularity issue*, our topic for §8.4.

8.4 The Granularity Issue

Given that a world represents such-and-such, what else must it represent? This is one of the most difficult issues surrounding impossible worlds. A logically possible world which represents that A and that B must thereby represent that $A \wedge B$, that $A \vee B$, that $\neg\neg A$, and so on. Which of these closure principles, if any, apply to impossible worlds? We call this the *granularity issue*. In this section, we'll discuss some considerations on granularity in general. We'll then look at how different considerations might apply to different hyperintensional concepts.

Let's begin with *Nolan's Principle* (Nolan 1997, 542):

(NP) If it is impossible that A, then there's an impossible world which represents that A.

Nolan thinks of this as a kind of unrestricted 'comprehension principle' for impossibilities. If true, it tells us something about both the nature and the scope of impossible worlds.

Given (NP), you might think we can show that impossible worlds needn't obey any particular logical closure principle (other than *identity*, $A \vDash A$). If that's right, then impossible worlds are *open worlds* (Priest 2005, Rantala 1982a). We defined open worlds in §5.3 as worlds which, in general, obey no logical closure principle other than $A \vDash A$. (From now on, when we say 'closed under no principle', we'll always mean 'no principle except $A \vDash A$'.)

The argument from (NP) to open worlds goes as follows. Take any putative closure principle, taking us from some premises $A_1, A_2, \ldots$ to C. If the principle isn't valid, then there's a logically possible world such that $A_1, A_2, \ldots$ but not C. If it is valid, by contrast, then it cannot be the the case that $A_1, A_2, \ldots$ but not C. But then, by (NP), there is an impossible world such that $A_1, A_2, \ldots$ but not C. Generalizing, for any closure principle, there is a world not closed under that principle.

This argument is flawed, for it misapplies Nolan's Principle. As stated, Nolan's Principle allows us to take any single sentence 'A'

which cannot be true, and infer the existence of an impossible world which represents that *A*. But the argument above considered multiple sentences, 'A_1', 'A_2', … and '*C*'. These cannot together be substituted into (NP) as stated. (We'll discuss more general principles, which Nolan may have had in mind, later in this section.)

We can always substitute for '*A*' in (NP) the single sentence, '$A_1 \wedge A_2 \wedge \cdots \wedge \neg C$'. This is impossible iff it cannot be the case that $A_1, A_2, \ldots$ but not *C*. So, using this sentence, (NP) legitimately gives us an impossible world w which represents that $A_1 \wedge A_2 \wedge \cdots \wedge \neg C$. However, there is no guarantee that w represents each A_i but does not represent that *C*. We haven't yet fixed which principles hold of impossible worlds and so, in particular, we can't assume those principle. It may be that w represents a conjunction without representing each of the conjuncts. It may represent both $\neg C$ and *C*.

We could instead try substituting a sentence of the metalanguage for '*A*' in (NP), such as this one: ''A_1', 'A_2', … are true, but '*C*' is not true'. Then (NP) gives us an impossible world w where that entire sentence is true. But again, this need not be a world where the A_is hold but *C* does not. To make that inference, w would need to support the inferences from *A*'s being true to *A*'s being the case and from *A*'s not being true to *A*'s not being the case. But since w is an impossible world, we can't assume that those principles hold of it. So however we try, we can't get directly from (NP) to the failure of all closure principles.

Priest (2016a) adopts two principles that are similar to, but stronger than, (NP): 'everything holds at some worlds, and everything fails at some worlds' (Priest 2016a, 5) and, for any distinct *A*, *B*, 'there are worlds where *A* holds and *B* fails' (Priest 2016a, 7). More specifically, in our terminology:

(8.5) For any *A*, there is a world which represents that *A* and a world which does not represent that *A*.

(8.6) For any distinct *A* and *B*, there is a world which represents that *A* but does not represent that *B*.

Priest calls these the 'primary directive' and 'secondary directive' on impossible worlds, respectively. The latter implies the former, which in turn implies (NP), but neither converse holds. To see this, first suppose that it is impossible that A. Then (8.5) says there is a world such that A, which by definition is an impossible world. So (8.5) implies (NP). Suppose instead that A is necessary. Then (8.5) implies that there is an impossible world which does not represent that A. But (NP) does not imply this. Since A is necessary, $\neg A$ is impossible, and so (NP) says there is an impossible world which represents that $\neg A$. But this need not be a world which does not represent that A: it may represent *both* that A and that $\neg A$, as some of the FDE-worlds from §5.4 do. In general, (NP) tells us only about the existence of worlds which represent such-and-such. It does not tells us about the existence of worlds which *fail* to represent such-and-such. That's why (8.5) is strictly stronger than (NP).

It's clear that (8.6) implies (8.5), but the converse does not hold. For any pair of sentences 'A' and 'B', (8.5) tells us that there is a world which represents that A and a world which fails to represent that B. But these could be distinct worlds. Since there's no way to infer that they're the same world, as (8.6) requires, (8.5) doesn't imply (8.6). So we get increasingly strong principles as we go from (NP) to (8.5) to (8.6).

To illustrate the extra power (8.6) gives us (over (8.5) and (NP)), consider Simplification, the inference from $A \wedge B$ to A, or Disjunction Introduction, from A to $A \vee B$. (8.6) directly entails that there are worlds where these rules fail. These are not FDE-worlds (which respect the classical rules for conjunction and disjunction).

But even (8.6) does not get us to the existence of open worlds, where any logical rule (except $A \vDash A$) can fail. For consider *adjunction*, from A and B to $A \wedge B$. Given (8.6), we may infer that there's a world which represents that A but not $A \wedge B$, and that there's a world which represents that B but not $A \wedge B$. But we can't get to a world which represents *both* that A and that B, but not that $A \wedge B$. (It clearly won't help to consider the premises conjunctively: there's no world which represents and does not represent that $A \wedge B$.)

To infer the existence of open worlds from such-and-such being impossible, we need a stronger principle still, such as the following:

(NP$^+$) If it is impossible that $A_1, A_2, \ldots$ but not C, then there's an impossible world which represents that $A_1, A_2, \ldots$ but not C.

This is very much in the spirit of the original (NP). Given it, for any logical principle (other that $A \vDash A$), there's a world which breaks that principle. If (NP$^+$) is true, then we need to include open worlds in our theory. However, there is virtually no gap between (NP$^+$) and that conclusion: (NP$^+$) says, in effect, that for any logical rule, there's a world which breaks that rule. It too needs an argument. And whereas Nolan's original principle (NP) has plenty of intuitive force (for those who believe in the existence of impossible worlds to begin with), we can hardly claim that for (NP$^+$).

Despite the failure of this argument, we believe that there exist open worlds. There are three arguments we can offer. The first is an argument from logical rules. If there are impossible worlds at all, then there are worlds which break *some* logical rule. More carefully: if there are qualitatively distinct impossible worlds, then there are impossible worlds which do not represent that A for every A. Such worlds are not governed by all the usual logical inference rules. Let's fix on the standard introduction and elimination rules for each connective, such as Adjunction and Simplification for conjunction. Each connective's meaning is intimately related to its introduction and elimination rules. (Just think about how Adjunction and Simplification fix what '$\wedge$' means.) But if one of these meaning-fixing rules can be broken by an impossible world, then surely any of them can. And if our domain of worlds does not in general respect any logical rule, we have a domain of open worlds. This argument doesn't give us a positive reason for thinking that all rules are broken by impossible worlds. Rather, it looks to convince us that there's no reason to reject that approach.

The second argument is an argument from epistemic states such as belief. (Jago (2014a) and Priest (2016a) give a similar argument.)

Take any putative closure principle P, from $A_1, A_2, \ldots$ to (distinct) C. If all worlds are closed under P, and we analyse belief in terms of accessible worlds, then any agent who believes that $A_1, A_2, \ldots$ will thereby be modelled as believing that C, too. But it is at least possible for some agent to believe $A_1, A_2, \ldots$ but not believe that C. So not all worlds are closed under P.

The argument is quite general. If some possible agent can believe the premises of some logical principle but not the conclusion, then there must be an epistemically accessible world which represents those premises but not the conclusion. That was basically our argument, in §5.4, against using a weaker-than-classical logic to account for epistemic states. If we restrict the worlds accessible to an agent to, say, the FDE worlds, then the agent remains logically omniscient, with respect to the first-degree entailments of her beliefs. To avoid that situation, for any closure principle, there must be an accessible world which doesn't obey that principle. That's the argument for open worlds from epistemic states.

The third argument for open worlds is an argument from counterpossible reasoning. (Priest (2016a) gives a similar argument.) Suppose we're debating whether some putative logical principle should be accepted. We might be studying philosophy of logic, and discussing arguments in favour of intuitionism or paraconsistency, for example. Then we might make suppositions about the validity (or otherwise) of Excluded Middle ($A \vee \neg A$), Double-Negation Elimination ($\neg\neg A \vDash A$), or the Explosion Principle ($A, \neg A \vDash B$). We'll consider what happens if that principle fails. If the principle is in fact valid, then we're making a counterpossible supposition (§1.3).

If we want to analyse counterpossibles using impossible worlds (as we propose in Chapter 12), then we need impossible worlds in our account which violate that particular logical principle. It seems that we can engage in this kind of supposition for *any* logical principle (other than $A \vDash A$). For each such principle, we'll need impossible worlds in our account which violate it. So impossible worlds in general cannot be closed under any logical principle (other than $A \vDash A$). That's just to say that we need open worlds in our account.

It doesn't follow from these arguments that one *single* world breaks all closure principles. It might be that each closure principle is broken by some world, although no world breaks them all. So, although the arguments above establish that we need open worlds in our account, it doesn't establish that the full range of open worlds (including worlds which respect no closure principle whatsoever) is required. Nevertheless, the simplest semantics available (given that open worlds are required at all) is to allow models to be built from any collection of open worlds, with no restrictions. Then some models will involve only worlds which meet a certain condition; but there will also be models containing worlds which don't meet that (or any other) condition.

For this perspective to make sense, we need a very fine-grained notion of how worlds represent. We need world-representation to be at least as fine-grained as the sentences of the object language. For otherwise, there would be distinct object-language sentences A, B such that a world represents that A iff it represents that B. But this gives us a closure principle ($A \dashv\models B$) which all worlds must obey, contrary to what we established above.

Linguistic ersatzism is the most straightforward way to realize this very fine-grained notion of world-representation. If impossible worlds are themselves sets of sentences, and any such set counts as a world, then world-representation will be as fine-grained as the worldmaking language itself. That is, for any pair A, B of worldmaking sentences, there will be a world which represents that A but not B. Moreover, as we argued in §3.6, linguistic ersatzism is probably the best way to make sense of impossible worlds. (This argument is neutral between linguist ersatzism proper and the linguistic variant of the hybrid account from §2.5, on which the possible worlds are Lewisian and the impossible worlds are sets of sentences.)

If we then make use of the class of all such worlds, we obtain extremely fine-grained analyses of hyperintensional concepts. But this is a problem in itself. We saw in §5.3, in our discussion of logical omniscience in epistemic logic, that this 'anything goes' approach to knowledge or belief tells us nothing about those concepts. If an agent can believe anything, whilst disbelieving anything else, then in

what sense do we have a model of belief at all? This 'anything goes' approach to belief seems equivalent to the purely syntactic approaches of Eberle (1974) and Moore and Hendrix (1979). These contain an arbitrary set of sentences, which are taken as representations of the agent's beliefs. If we can do things that way, why bother with the worlds apparatus in the first place?

Here, we have reached a deep puzzle about hyperintensionality. Hyperintensional concepts require our models to be extremely fine-grained. Yet when we achieve that fine granularity, we seem to have surrendered the benefits of the approach.

We find the arguments above, over the granularity of worlds, convincing. So we accept the open worlds picture: in general, worlds need be closed under no particular logical rule. Our metaphysics of what worlds are, and of how they represent, must be in accord with this principle. But it doesn't follow that 'anything goes' with any analysis of a hyperintensional concept, for two reasons. First, a particular concept may require a certain kind of world, obeying certain conditions. When analysing possibility and necessity, for example, we restrict our attention to the possible worlds only. Although these are not intensional concepts, a similar restriction applies to many hyperintensional concepts. If we're interested in what's possible or necessary *with respect to intuitionistic logic*, or *with respect to paraconsistent logic*, we restrict our attention to the worlds which obey those logics.

There are (hyperintensional) notions of *semantic content* which, we think, require us to restrict the domain of worlds. (We discuss one such notion in §9.6.) These notions require worlds more fine-grained than classical possible worlds, but not as fine-grained as open worlds. We obtain those worlds by narrowing down the class of all open worlds, based on certain principles we want to enforce on our analysis. None of these cases allow that 'anything goes'. As we shall see in §9.6, there are substantial, non-trivial equivalences on semantic contents, which we can capture in our impossible worlds framework. That's the first reason why adopting open worlds doesn't imply 'anything goes' for a hyperintensional concept.

The second reason is that an account of a hyperintensional concept may involve non-trivial structure, even if its worlds are themselves unstructured open worlds. One example is given by worlds-theories of counterfactuals (§1.2), which involve a similarity metric on worlds. We can adopt that approach even if we include open worlds in the account. (Again, we discuss counterpossible conditionals in detail in Chapter 12.) The metric can give us structure when we want it, but not when we don't. Let's counterfactually suppose that intuitionistic logic is the correct account of validity, as in our example above. Then the similarity metric will select worlds where intuitionistic principles operate but which (let's assume) are otherwise most similar to our own. This approach tells us that, if intuitionistic logic were correct, Double Negation Elimination would not be valid. But it also tells us that, if intuitionistic logic were correct, Nottingham would still be wet today. And it doesn't tell us that, if intuitionistic logic were correct, there would be flying elephants in Amsterdam. (This all depends on the details of the similarity metric.)

In this kind of approach, anything may be supposed. But it isn't the case that anything goes within a supposition. The similarity metric (or some other metric, if you prefer) gives us appropriate structure within any counterfactual supposition. There may or may not be principles governing counterfactual supposition across the board (Chapter 12). But even if there aren't, it's not the case that 'anything goes' within any counterfactual supposition. That's enough to convince us that the open worlds approach isn't trivial.

What of the worry, mentioned previously, that a model of belief with open worlds is trivial? This is one of the toughest problems to address, given that there seem to be no necessary closure conditions on what we can believe. Our answer (much as in the case of counterpossibles) is that a *model* may include important structure, even if the worlds it involves do not. We discuss the issue in §§10.3–10.4 and present a formal model in §10.5, based on some ideas discussed in §9.5.

There is one further worry the open worlds approach should address: that it results in disjunctive truth-conditions and implausible

consequences for a theory of meaning. (We touched on this issue in §4.3.) We'll discuss the worry and offer a response in §8.5.

8.5 The Compositionality Objection

Compositionality is the principle that the meaning of a complex expression is a function of the meanings of its constituent expressions. It's commonly taken to be a mandatory feature of any adequate theory of meaning. The argument is that, as competent speakers of a language, we are in principle capable of grasping the meanings of a potentially infinite number of sentences. And since we've learnt the meanings of a limited number of words, this is possible only if the meanings of complex sentences are obtainable recursively from the meanings of their constituent parts (Davidson 1965).

Take this very simple example. If you know what 'badgers are great' and 'and' mean in English, then you know what

(8.7) Badgers are great, and badgers are great, and badgers are great, and badgers are great, and badgers are great, and badgers are great

means. Chances are, you never encountered that sentence before reading it here, yet you understood it straight away. How so? Because you understood all the component parts, of which the meaning of the whole is a function. That's compositionality.

The argument applies equally to notions of content: the content of a complex expression must be a function of the content of its constituents. The content of a conjunction $A \wedge B$, for example, must be a function of the contents of A and of B. The content of $A \vee B$ will be some other function, also of the contents of A and of B.

The possible worlds account of propositions (§8.3) clearly meets this requirement: if the contents of A and of B are sets of possible worlds, then the content of $A \wedge B$ is their intersection and the content of $A \vee B$ is their union. Since each pair of sets has a unique intersection

and union, this account of content (for the Boolean connectives, at least) is compositional.

In this example, we see a link between compositionality and truth-conditions. Given that $\wedge$ requires the truth of both its conjunctions, the possible-worlds content of $A \wedge B$ must be all those worlds where both A and B are true. On any account of worlds, that's just the intersection of the A-worlds with the B-worlds. So here, there is a direct link between the connective's truth-condition and the function on contents. Similarly for the other truth-functional connectives.

This link breaks down with open worlds. At an open world, it is in general not the case that $A \wedge B$ is true just in case both A and B are true. There are open worlds where badgers are great but our badger sentence (8.7) is not true. So, when contents may include open worlds, we do not obtain the content of $A \wedge B$ by taking the intersection of the A-worlds with the B-worlds. Indeed, insofar as the truth of $A \wedge B$ is independent of the truth of A and of B at an open world, there appears to be no function from the open-world contents of A and B to the content of $A \wedge B$. But that's just to say that the account isn't compositional.

This is a serious worry for the open worlds approach. If it can't be met, then it could well be a fatal objection. (Note that it should be seen as a worry concerning meaning or content, not validity. For we analyse validity in terms of what's the case according to the *possible* worlds; and for possible worlds, the usual recursive clauses for the connectives hold.)

Our response to the problem is in two parts. First, we distinguish two aims for a formal theory. A formal theory might be (part of) a theory of meaning, in which case, it must be compositional. But a formal theory may be intended as a useful model of some notion (information, or belief, or logical consequence), without claiming to be a theory of meaning. In this sense, a model might get things extensionally correct, whilst not respecting the underlying mechanisms of *why* the modelled concept works they way it does. In short, the accounts we've presented might be fine for some tasks, but not for others, including giving a theory of meaning. (This kind

of attitude seems implicit in the work of those who have employed non-normal worlds for various logical purposes, which we examined in Chapters 4–7.)

The second, and less defensive, part of our response is to show that at least some open worlds-based approaches *are* compositional. This would show that a full theory of meaning can be given by including open worlds. To do this, we'll need to return to the metaphysics of worlds, which we discussed in Chapters 2 and 3.

Consider the various metaphysics of worlds we gave throughout those chapters, and what they say about truth-at-a-world. If w is a concrete world, then A is true at w iff things are such that A, once quantifiers are restricted to w. If w is an ersatz world, then truth-at-w might be a matter of which properties comprise (or are encoded by) w. Or it might be a matter of which propositions or worldmaking sentences are members of w. Or it might be some other feature of w's construction.

We're going to focus on the linguistic ersatz approach, on which worlds are sets of sentences (§3.6). Such worlds represent *explicitly*, by containing a sentence that says such-and-such. These need not be sentences of the object language. Indeed, it's best they're not, else we make little progress in connecting the object language to reality. Let's write 'A^*' for the translation into the worldmaking language of the object language sentence 'A'. Truth-at-a-world is then defined in terms of world-membership:

(T_w) A is true at w iff $A^* \in w$

This definition applies to all worlds w and all sentences A, without regard for whether w is possible or impossible, open or not, and whether A is atomic or complex.

Next, let '$[\![A]\!]$' denote the content assigned to the object language sentence A: the set of all worlds at which A is true. This is the set of all worlds which contain A^*, $\{w \mid A^* \in w, w \in W\}$ (where W is the set of all worlds). But, given that any set of worldmaking sentences is a world in W, if we add A^* to every world in W, the set we obtain,

$\{w \cup \{A^*\} \mid w \in W\}$, is none other than $[\![A]\!]$. Moreover, these worlds all have A^*, and no sentence but A^*, in common. More precisely, their intersection, $\bigcap [\![A]\!]$, is none other than $\{A^*\}$.

Given these facts, we can always move back-and-forth between the content c of an object language sentence A and the corresponding worldmaking sentence A^*. So, as long as we can build up complex worldmaking sentences from more basic ones, in roughly the way we do for object language sentences, the semantics will be fully compositional. What follows is one way to make this precise.

Let's abstract from the precise grammatical details and suppose that complex worldmaking sentences are built using 1-place sentence operators, written prefix, and 2-place sentence operators, written infix. (These might include the Boolean connectives, various conditionals, and unary and binary modalities). Let O_1 be any such 1-place operator and O_2 be any such 2-place operator, and let $x^\frown y$ be the concatenation of worldmaking strings x and y. Then $O_1{}^\frown A^*$ and $A_1^*{}^\frown O_2{}^\frown A_2^*$ are worldmaking sentences.

Next, we define two semantic functions, f_1 (on the contents of a 1-place operator and a sentence, both from the object language) and f_2 (on the contents of a 2-place operator and two sentences):

$$f_1(o, c) = \left\{ w \cup \{o^\frown x \mid x \in \bigcap c\} \mid w \in W \right\}$$

$$f_2(o, c_1, c_2) = \left\{ w \cup \{x^\frown o^\frown y \mid x \in \bigcap c_1, y \in \bigcap c_2\} \mid w \in W \right\}$$

To understand what's going on here, recall that, when c is the content of an object language sentence, $\bigcap c$ will be some singleton $\{A^*\}$. So the definition of f_1 sets x to be equal to some worldmaking sentence A^*, to which it concatenates o, and then returns the set of all worlds containing that complex worldmaking sentence. Let's write '$\neg^*$' for the translation into the worldmaking language of the object language negation '$\neg$'. Then $f_1(\neg^*, [\![A]\!])$ returns $[\![\neg A]\!]$, the set of worlds according to which $\neg A$ is true.

The same goes for f_2. Given an object language conjunction $A \wedge B$, we use f_2 on $\wedge^*$ (the worldmaking translation of '$\wedge$') and the contents of A and B. Then $f_2(\wedge^*, [\![A]\!], [\![B]\!])$ returns $[\![A \wedge B]\!]$, the content of

$A \wedge B$. That's all we need to demonstrate that the semantics is fully compositional.

How is it that this approach is compositional, when various semantics with open worlds presented throughout Chapters 4–7 are not? We gave those semantics in terms of a primitive valuation function or relation, assigning truth-values to sentences at worlds. For open worlds, the valuation (or relation) assigns truth-values to all sentences directly, and without restriction. That valuation is non-recursive. But the content assigned to a sentence depends on that valuation, and so this notion of content is non-recursive. Compositionality in the linguistic ersatz approach, by contrast, relies on a recursive translation from object language sentences to worldmaking sentences. The worldmaking sentence A^* does a lot of the heavy lifting in giving A's content. That's why the linguistic ersatz approach delivers a recursive characterization of the content of any object language sentence.

Chapter Summary

We began our discussion by asking whether hyperintensionality is a genuine phenomenon, or rather, a feature to be explained away (§8.1). We then focused on the epistemic case, considering arguments from Stalnaker and Lewis which attempt to explain away hyperintensionality (§8.2). We argued that they are not successful. We then considered the argument for a genuinely hyperintensional notion of content (§8.3). Having made the case for genuine hyperintensionality, we turned to the *granularity issue* (§8.4): how fine-grained are impossible worlds? This is one of the most difficult issues any theory of hyperintensionality faces. We then returned to the *compositionality objection*, and argued that some accounts of impossible worlds deliver a fully compositional theory of meaning (§8.5).

9

Information and Content

9.1 Informative Statements

Here are some informative statements. It's currently cold but dry in Nottingham. James Newell Osterberg is Iggy Pop. Fermat's Last Theorem is true. If all truths are knowable, then all truths are known. And here, by contrast, are some uninformative statements. No bachelor is married. Iggy Pop is Iggy Pop. $1 + 1 = 2$. Either all truths are knowable or they're not.

The first list of statements might be informative to some people and not to others. For someone standing outside in Nottingham right now, it's probably not informative to be told that it's currently cold but dry there. For someone who knows lots about Iggy Pop, it's probably not informative to be told that James Newell Osterberg is Iggy Pop. To those who've encountered Fitch's paradox, it's probably not informative to be told that, if all truths are knowable, then all truths are known. To modern mathematicians, it's probably not informative to be told that Fermat's Last Theorem is true.

Whether a statement is informative to someone depends on what information that person already has. It also depends on the way in which they have that information. Take our person standing outside in Nottingham right now. She may be confused about where she is. If she doesn't know she's in Nottingham, then experiencing the weather in her immediate surroundings won't help her to conclude that it's currently cold and dry in Nottingham. Then her weather app, which says *currently cold and dry in Nottingham*, might be informative to

her. In some sense, one might think, she already knew this. She knew it's currently cold and dry where she is, and Nottingham is where she is. But still, it was informative to her to be told that it's currently cold and dry in Nottingham. What's informative depends also on how the information is presented.

In this chapter, we investigate accounts of what it is for a statement to be informative. We're interested in what information is, in and of itself. We're also interested in how a statement gets to be informative. As the examples above suggest, logical and mathematical truths can be informative (to some people, at least). This in turn hints that an impossible worlds framework is a promising way to understand this phenomenon. Yet, as the examples also suggest, not all logical and mathematical truths are informative (to anyone). So, if we go down the impossible worlds route, we need to be careful about *which* impossible words are included in the analysis.

Similarly, some but not all identity statements are informative (to some people). If identity is necessary, as most philosophers hold after Kripke (1980), then impossible worlds seem to be an attractive way to go. But again, we will need to be careful about which impossible worlds feature in the analysis. Informativeness, it will turn out, is a very puzzling concept.

9.2 Information as Ruling Out Scenarios

According to a popular analysis, for a sentence (or the expressed proposition) to be informative is for it to rule out certain scenarios, or would-be possibilities. Hintikka (1962) gave the classic presentation of this view, which Chalmers (2002a, 2010), Lewis (1975, 1986b), and Stalnaker (1976b, 1984) then put to work in various ways. Van Benthem (2011) and Van Benthem and Martinez (2008) discuss the recent literature.

The proposition *that it often rains in Manchester* is informative because it excludes scenarios in which it rains infrequently in Manchester. Before an agent comes to believe that proposition, it was

possible, as far as she was concerned, that it rains infrequently in Manchester. In coming to believe that proposition, she ceases to treat such scenarios as ways the world might be, for all she knows.

We can think of all scenarios according to which it often rains in Manchester as constituting a notion of content for 'it often rains in Manchester' which is suitable for various epistemic purposes. To believe that proposition is to treat *only* those scenarios according to which it often rains in Manchester as being doxastic possibilities; and to know that proposition is to treat *only* those scenarios according to which it often rains in Manchester as being epistemic possibilities (Chapter 5). That content is informative for an agent iff coming to believe (or know) that proposition narrows down her doxastically (or epistemically) accessible scenarios. To be informative at all, therefore, a statement must have a non-empty content.

If the scenarios in question are all possible worlds, then problems ensue. In §1.2, we introduced the Bar-Hillel-Carnap theory of information (Bar-Hillel and Carnap 1953). This claims that the informative role of a sentence consists in splitting the totality of possible worlds into those where it is true, and those where it is false. The consequence outlined in §1.3 is that identity statements of the form $a = b$, and metaphysical, logical, and mathematical truths, all end up being treated as uninformative.

As a consequence, the Bar-Hillel-Carnap theory denies that any logical deduction or mathematical proof can ever be informative. This is implausible as a characterization of information for limited, resource-bound, and fallible cognitive agents like us all. The issue, which Floridi (2015, §4.1) calls the 'Bar Hillel-Carnap paradox', is in fact just a variation on the logical omniscience problem in epistemic logic (Chapter 5), which we'll discuss further in Chapter 10.

Acknowledging the problem, Bar-Hillel and Carnap say that their account should not

> be understood as implying that there is no good sense of 'amount of information' in which the amount of information of these sentences will not be zero at all, and for some people, might even be rather high. (Bar-Hillel and Carnap 1953, 229)

This 'good sense' is a kind of 'psychological information' (1953, 229), on which the Bar-Hillel-Carnap theory has nothing further to say. On this view, there is one kind of information, semantic (and, presumably, non-psychological) information, which applies to contingent, empirical statements. And there is distinct 'psychological' notion which, in some unspecified way, makes sense of the informativeness of certain necessary truths.

That approach looks messy and poorly motivated. We don't find pre-theoretical reasons for wanting to divide notions of information along these lines. Learning about the logical consequences of some supposition, or of a putative move in a game of chess, seems to us to be informative, in much the same sense that contingent statements can be informative. If asked whether it's currently raining in Sydney, or whether Qe4 is a good next move, an uninformed agent might in each case say 'could be'. In each case, becoming informed rules out would-be possibilities for the agent in question. Both cases link to belief and knowledge in the same way. In each, gaining information leads to new beliefs and, in the right circumstances, to fresh knowledge. Moreover, both cases have a psychological element. Both can be surprising; both contents can be the objects of an agent's hopes or fears; both may interact with the agent's emotions. In short, we think a unified notion of semantic informational content, one which can deal with both cases, is preferable.

We don't thereby want to claim that there is only one good notion of information. Far from it: there are many such notions, and more than one may be theoretically useful. In §9.1, we contrasted the concept of being *potentially informative to some agent* with *being informative to agent x*, for a specified *x*. In §9.6, we'll discuss a distinct notion of informational content, which concerns *what is said* by a speaker in making an utterance. Our overall strategy is the one we discussed in §8.4. We start with fine-grained worlds. We impose additional *inter*-world structure, or *intra*-world closure conditions, depending on the concept of information under investigation.

9.3 Informative Identities

James Newell Osterberg recently turned 70. That information would be of little interest to us (to Franz and Mark, at least) if we didn't also know that James Newell Osterberg is Iggy Pop. Given what we know about his time in The Stooges, it's somewhat surprising that he's made it to 70. If we didn't know that James Newell Osterberg is Iggy Pop, we'd not find it so surprising that James Newell Osterberg has reached 70. 'James Newell Osterberg is Iggy Pop' is an *informative identity*. It allows us to connect our attitudes to particular bits of information to other bits of information. It allows us to connect our *Iggy Pop* information, and the attitudes we take to it, to our *James Newell Osterberg* information.

How can 'James Newell Osterberg is Iggy Pop' be informative? Since it's true, 'Iggy Pop' is a name for Osterberg, as is 'James Newell Osterberg'. Both names pick out the same individual. So, semantically, 'James Newell Osterberg is Iggy Pop' would seem to express the proposition *that Osterberg is Osterberg*. But that's utterly trivial and uninformative. So how are we to understand the information conveyed by 'James Newell Osterberg is Iggy Pop'?

This is a version of *Frege's puzzle* (Frege 1892). Frege's own solution was (in effect) that propositions are not identified by the worldly entities they are about, but rather by the *modes of presentation* of those worldly entities (he called them *Sinne*, senses). If the modes of presentation of '*a*' and '*b*' differ, then the Fregean propositions (which Frege called 'thoughts') *that Fa* and *that Fb* differ. This allows an agent to stand in a relation (such as *believing* or *knowing*) to one proposition but not the other. So, in particular, one can learn that Iggy Pop has turned 70, even if one already knew that Osterberg has turned 70. In this way, Frege can explain how replacing 'Osterberg' with 'Iggy Pop' can be informative. In particular, he can explain how 'Osterberg is Iggy Pop' is informative, whereas neither 'Osterberg is Osterberg' nor 'Iggy Pop is Iggy Pop' are.

To support this approach, Frege requires some rather elaborate semantic machinery. On a straightforward semantic account, '*a*' refers

to *a*, '*b*' to *b*, and it is these referents, *a* and *b*, which feature in the truth-conditions for the utterance in question. And so it goes on Frege's analysis, in *direct* discourse. But in *indirect* discourse, such as the sentence following 'it is informative that' or 'believes that', '*a*' and '*b*' refer not to *a* and *b* but to their mode of presentation. It is these modes of presentation which feature in the truth-conditions of the utterance as a whole. Thus, on the Fregean view, indirect contexts induce a switch of reference, from the usual worldly entities to their modes of presentation.

The Fregean approach is rich and powerful. It's not our intention here to evaluate it. Our topic is impossible worlds. Our aim is to show that a worlds-based approach can analyse informative identities (and belief reports: see Chapter 10) as well as the Fregean approach.

If we're successful, then we can avoid a question that troubles the Fregean approach: just what are senses? Frege speaks of the 'realm of senses', distinguished from the 'realm of reference' (Frege 1956). It seems that, for Frege, senses are *sui generis* entities, neither physical nor mental (see Dummett 1993a, 154). Dummett notes that, for Frege, 'the realm of sense is a very special region of reality; its denizens are, so to speak, things of a very special sort' (Dummett 1993a, 154). Senses must be entities of some kind, else they could not be referents in belief-contexts (as Frege's theory claims). But if they are primitive, non-causal abstract entities, how can we refer to them in belief reports?

There may well be good answers to these questions. (Chalmers (2002c) identifies Fregean senses with 'primary' or 'epistemic' intensions, a particular function from possibilities to extensions. But Chalmers's approach makes no concession to hyperintensional notions.) We avoid the worry entirely if we can show that a worlds-based approach is up to the job. (Indeed, Bjerring and Rasmussen (2017) and Jago (2014a) suggest that the best approach to understanding senses is in terms of functions on possible and impossible worlds.)

Our alternative to the Fregean approach goes as follows. If we accept Nolan's Principle (NP), or any of the other principles from §8.4, then for any non-empty terms '*a*' and '*b*', there are guaranteed

to be worlds which represent that $a = b$ and worlds which represent that $a \neq b$. For suppose $a = b$. Then the actual world represents that $a = b$. And since it is impossible that $a \neq b$, by (NP) some impossible world represents that $a \neq b$. If we suppose instead that $a \neq b$, then the argument is similar. (Just *how* a world represents that $a \neq b$, given that in fact $a = b$, is a further question. Jago 2014a, §§5.5–5.6 is one attempt at a solution.)

Since there are guaranteed to be worlds according to which $a \neq b$, a true identity statement '$a = b$' is guaranteed to have a non-empty content. An agent may take any doxastic attitude to that content: belief, disbelief, or neither. If she does not believe it, it is because she takes some of the worlds where a is not b (which are, in fact, impossible) to be ways the world could be, for all she knows. This is compatible with her being a rational agent and perfectly competent language-user (she might even be a heavily idealized agent who knows all a priori truths). So there is no rational compulsion for her to believe that content. She may go from not believing it to believing it and, in the right circumstances, she may gain it as knowledge in the process. If she does, it is informative to her.

On this approach, true identity statements '$a = b$' are potentially informative. This approach maintains the benefits of the Fregean approach, but without relying on reference switching mechanisms. (In Chapter 10, we'll also argue that the impossible worlds approach can make sense of belief ascriptions without resorting to reference switching.)

Won't the approach incorrectly treat '$a = a$' as being informative, too? The worry arises because $a = a$ is a logical truth, and hence it's logically impossible that $a \neq a$. So, given (NP) or one of the stronger principles from §8.4, there are worlds which represent that $a \neq a$. Mustn't we then treat '$a \neq a$' and being potentially informative?

We can resist this final move. Some contents are not suitable objects of epistemic attitudes. Not all impossible worlds are epistemically possible: some are not epistemically accessible for any agent. Some such worlds represent blatant contradictions, like representing some A as being both true and false. On the account of logical information

we offer in §9.5, such worlds are *deeply* epistemically impossible. No such world is accessible to any possible agent. The details of the view will have to wait until §9.5. But we can already anticipate that, in just the same way, worlds which represent some a as not being self-identical are deeply epistemically impossible. So, although there are sets of worlds which represent that $a \neq a$, no such set of worlds is an *epistemic* content.

To be informative, a statement 'A' must be capable of being disbelieved. It must be possible for an agent to believe that A, or to believe that $\neg A$ instead. But since $a \neq a$ cannot be believed (on the account we're suggesting), it follows that $a = a$ is treated correctly as an uninformative identity.

9.4 Informative Inference

This section draws on Jago (2013b). Deductive reasoning is essential to philosophy, mathematics, and logic. In those areas and others, its use is beyond question, and this must be so, at least in part, because of information it conveys. But *how* can deduction carry information, if, in some sense, the premises already guarantee the conclusion? In 'The Justification of Deduction', Dummett (1978a, 297) asks how deduction can be both justified and useful. If it is justified, it must be guaranteed to preserve truth from premises to conclusion. To be useful, it must inform us of something.

How, wonders Dummett, can the move from premises to conclusion be informative, if the former already guarantee the latter? It is 'a delicate matter so to describe the connection between premisses and conclusion as to display clearly the way in which both requirements [justification and usefulness] are fulfilled' (Dummett 1978a, 297). The task is to capture this notion of information content whilst respecting the fact that the content of the premises, if true, already secures the truth of the conclusion.

We might think of the information content of a valid deduction $\Gamma \vdash A$, from premises Γ to conclusion A, in terms of the differences

an agent's belief state might undergo in performing that deduction. We can consider an agent who initially believes the premises but not the conclusion, and who ends up believing the conclusion (on the basis of the deduction she performs). Alternatively, we can think in terms of an agent discovering the incompatibility of the premises Γ with the conclusion's falsity. Either way, we are analysing some relationship between the content of the premises and the content of the conclusion.

Let's use the notion '$|A|$' to denote the set of worlds (possible or impossible) which represent that A. For sets of sentences Γ, we'll use '$|\Gamma|$' to denote the set of worlds which represent that B, for each $B \in \Gamma$. Our approach throughout this book has been to analyse notions of content in terms of possible and impossible worlds. In our present setting, the worlds in question have to outrun the logically possible ones. For suppose we limit each set $|A|$ to the possible worlds. Then if Γ entails A, $|\Gamma|$ already includes $|A|$, and so already excludes A's being false. So, however we analyse the relationship between premise and conclusion contents, we will be working with sets that include logically impossible worlds. As a minimal requirement, what they represent must not be closed under classical logical consequence.

A popular place to look for such worlds is the model theory of *paraconsistent* and *paracomplete* logics, which we encountered in §5.4 in the guise of *FDE worlds*. At an FDE world, a sentence may be true, false, both, or neither. This is accomplished by replacing the usual valuation function with a relation, ρ, which may relate a sentence to 1 or 0, to both, or to neither.

Let's take the content of premises and conclusion to be given in terms of such worlds. Our first notion of content of an inference $\Gamma \vdash A$ focuses on the difference between the premises without the conclusion and the premises with the conclusion. This amounts to those worlds according to which the premises are true, but the conclusion is not: $|\Gamma| - |A|$. Call this $content_1$. Our second notion analyses the content of $\Gamma \vdash A$ in terms of those worlds where the premises are true but the conclusion is false. In FDE worlds, A is false iff $\neg A$ is true, and so this notion of content amounts to $|\Gamma| \cap |\neg A|$. Call this notion $content_2$.

These notions are classically equivalent but differ in our paraconsistent and paracomplete FDE setting, since, as we saw in §5.4, being false and failing to be true come apart at FDE worlds. We shall say that an inference is *trivial*$_1$ (or *trivial*$_2$) just in case its content$_1$ (or content$_2$) is the empty set. We'll use 'non-trivial$_{1/2}$' and 'contentful$_{1/2}$' interchangeably, and we'll reserve 'trivial', without a subscript, to capture the non-technical sense in which inferences like $A \vdash A$ (but not all valid inferences) seem obvious and uninformative.

FDE models are not in general closed under *modus ponens* for the material conditional $\supset$: there are worlds w where both $A \supset B$ and A are true, but B is not: $\rho_w(A \supset B)1$ and $\rho_w A1$ but not $\rho_w B1$. So $|\{A \supset B, A\}| - |B|$ is non-empty: *modus ponens* on $\supset$ is non-trivial$_1$. Similarly, it is non-trivial$_2$, since there are worlds w where both $A \supset B$ and A are true, but B is false: $\rho_w(A \supset B)1$, $\rho_w A1$ and $\rho_w B0$.

On this picture, not all valid deductions come out as being contentful$_1$. The deduction $A, B \vdash A \wedge B$ remains trivial$_1$, since any FDE world verifying A and B individually also verifies $A \wedge B$, and so $|\{A, B\}| - |A \wedge B|$ is empty. Indeed, any classically valid inference whose only connectives are '$\wedge$' and '$\vee$' will be deemed trivial$_1$, on this view. This is a puzzling feature for an account of content. *Modus ponens* and Conjunction Elimination (for example) do not seem to be wholly different kinds of inference rule. If one is deemed trivial, then why not the other?

By contrast, *every* valid inference is deemed contentful$_2$. In FDE worlds, the conclusion may be both true and false. So even where the premises guarantee the truth of the conclusion in our FDE setting, they do not thereby rule out its falsity. Even the most seemingly trivial inference of all, $A \vdash A$, is deemed contentful$_2$. Its content$_2$ is $|A| \cap |\neg A|$, which is the set of all worlds w where A is both true and false: $\rho_w A1$ and $\rho_w A0$. This is an even worse consequence than the results for content$_1$. Surely some inferences are so trivial as to contain no information whatsoever. If $A \vdash A$ is deemed informative, then we seem to have a worthless notion of information.

There is a deeper problem with the FDE worlds approach: it fails to explain why the worlds it provides are suitable tools for analysing

epistemic notions of content and information. It is a consequence of the account that both the content$_1$ and content$_2$ of a valid deduction $\Gamma \vdash A$ can contain only glutty worlds, which assign both 0 and 1 to some sentence. To see why, assume that $\Gamma \vdash A$. Then for any consistent assignment ρ_w on which $\rho_w B1$ for each $B \in \Gamma$, $\rho_w A1$ too, and hence (given consistency) not $\rho_w \neg A1$. But then $w \notin (|\Gamma| - |A|)$ and $w \notin (|\Gamma| \cap |\neg A|)$. So each notion of content can contain only explicitly contradictory worlds, at which some A is both true and false.

The problem is that it is hard to see why such explicitly contradictory worlds should play a role in an epistemic notion of content. If what a world represents is obviously impossible to any agent who meets minimal standards of rationality, then there's no sense in which ruling out that world corresponds to gaining new information.

This is the very feature which makes our problem difficult. If we are to model the content of a valid deduction as a set of worlds, then we have to admit impossible worlds. But obviously impossible worlds, representing explicit contradictions, cannot feature in any account of rational attitudes. And on the FDE-worlds account of deductive content, the obviously impossible worlds are *all* we're left with. In short, our question is difficult because it requires us to find worlds which are impossible, but not obviously so.

Our problem, therefore, is not merely to find worlds not closed under classical consequence. The problem is to provide a notion of a world which is logically impossible, but not obviously so. Lewis (in arguing against paraconsistent logic) puts the point nicely:

> I'm increasingly convinced that I can and do reason about impossible situations. … But I don't really understand how that works. Paraconsistent logic … allows (a limited amount of) reasoning about *blatantly* impossible situations. Whereas what I find myself doing is reasoning about *subtly* impossible situations, and rejecting suppositions that lead fairly to blatant impossibilities. (Lewis 2004, 176)

On Lewis's analysis, 'make-believedly possible impossibilities' might well have a use in the analysis of content, but:

> The trouble is that all these uses seem to require a distinction between the subtle ones and the blatant ones (very likely context-dependent, very likely a matter of degree) and that's just what I don't understand. (Lewis 2004, 177)

Hintikka (1975), whilst addressing the logical omniscience problem head-on, makes a similar point (see §5.3). He argues that, for epistemic purposes, impossible worlds must be 'subtly inconsistent' worlds which 'look possible but which contain hidden contradictions' (Hintikka 1975, 476–8). The core problem with FDE worlds (and will all similar approaches) is that they are either logically possible, or blatantly impossible.

How can we make sense of a world being subtly impossible? We present one attempt in the next section.

9.5 Vague Logical Information

Jago (2013b, 2014a) argues that we should view the problem of informative inference as an instance of the problem of vagueness. It seems that the deductive moves from $A \wedge B$ to A, or from $A \rightarrow B$ and A to B, are uninformative. All such moves seem utterly trivial. The problem then is that any deductive inference can be reconstructed by chaining together enough of these seemingly trivial inferences. If each step is trivial and uninformative, then we seem committed to saying that the entire deductive inference is trivial and uninformative. Yet some deductive inferences are not trivial, and can be informative. Something is amiss here. Dummett makes a similar point:

> When we contemplate the simplest basic forms of inference, the gap between recognising the truth of the premisses and recognising that of the conclusion seems infinitesimal; but, when we contemplate the wealth and complexity of number-theoretic theorems which, by chains of such inferences, can be proved … we are struck by the difficulty of establishing them and the surprises they yield. (Dummett 1978a, 297)

According to Jago (2013b), the problem has the structure of a sorites series. Suppose you've just marked 100 student essays and, as it happens, each got a different percentage mark from all the others. (So, each positive integer up to 100 is the grade of exactly one of the essays.) The top-marked ones were great. The lowest marked ones were pretty awful. But it's hard to say precisely which ones were good, which ones not good. Is it that all and only those with a mark over 40%, or 55%, or 68%, were the good ones? If so, what about the essay which scored 40% (or 55%, or 68%)? Was it so much worse that the essay which scored just 1% more? Surely not!

It seems absurd to pronounce for sure that only those essays scoring over (say) 55% were any good. Since there's no appreciable difference in quality between each percentage point, it seems that, if we judge any essay to be good, we should also judge the one scoring just 1% less to be good, too. But, as the 100% essay is clearly good, we are then at risk of judging, incorrectly, that all are good. The puzzle is to make sense of truth and inference in a vague language, so that not every essay is counted as being good.

Similarly, the task in the case of deduction is to make sense of a notion of content such that some, but not all, valid deductions are informative. And just as in the case of the essays, we have to do this without drawing an artificially sharp line between those deductions that are informative and those that are not. On this way of thinking about things, the normative notion of logical content is a vague notion, because chains of seemingly uninformative inferences can give rise to informative deductions.

Saying that the content of logical inferences may be indeterminate is not to provide a solution to these issues, however. It is merely to indicate that the problem has a certain form, one which we meet in other cases of vague predicates. Nor is this to say that we can pass the buck, by placing the problem of logical information at the feet of those working on theories vagueness in general. A philosophical theory of vagueness, as commonly understood, is a theory of how vagueness arises (is it metaphysical? semantic? epistemic?) together with an account of how vague predicates work. If we agree that

the language of logical information, content, and inference can be vague, then a full solution will certainly need to appeal to a general philosophical theory of vagueness. But a full solution to our present problem requires more than this.

A solution to our problem should consist in a model of logical information which explains why we find trivial inferences utterly uninformative, yet capable of being chained together into informative deductions. Let's consider further the analogy with more common cases of vagueness. In a deduction-sorites, each inference rule may be associated with a tolerance principle, saying that if such-and-such deduction is uninformative, then so is the one extended in such-and-such way.

If we have a trivial derivation of A from premises Γ, for example, then the tolerance principle associated with Disjunction Introduction says we also have a trivial derivation of $A \vee B$ from Γ. If we write the relationship of trivial derivation as '$\vdash_{\text{triv}}$', then this tolerance principle can be written:

$$\frac{\Gamma \vdash_{\text{triv}} A}{\Gamma \vdash_{\text{triv}} A \vee B}$$

There are similar tolerance principles for each connective, covering both appearances of the connective on the right-hand side (as the conclusion) and on the left-hand side (in the premises).

Together, these principles give us a proof system for $\vdash_{\text{triv}}$, which coincides with the underlying derivability relation. In other words, $\Gamma \vdash_{\text{triv}} A$ iff $\Gamma \vdash A$: all derivations are trivial! Since that's clearly wrong, at least some of these tolerance principles (expressed as proof rules) are incorrect. Logic dictates that a solution must reject the tolerance principle for at least one connective in each functionally complete set (such as $\{\neg, \wedge\}$, $\{\neg, \vee\}$, and $\{\rightarrow, \bot\}$). If we did not, we could infer that all derivations are uninformative. So one option may be to reject the tolerance principles for some (e.g., $\neg$ and $\rightarrow$) but not all connectives.

We claimed in Jago (2014b), however, that *all* of these tolerance principles should be rejected. The argument is that the inference

rules for '∧' and '∨' stand to the meaning of those concepts just as the inference rules for '→' stand to its meaning. (That is not to say that those rules constitute those meanings, but merely that there is a clear relationship between meaning and inference rules.) So, if the meaning of '→' does not guarantee that uses of *modus ponens* are uninformative, then neither can the meanings of '∧' and '∨' guarantee that inferences involving '∧' and '∨' are uninformative. But what could guarantee that a given kind of inference is always uninformative, if not the meanings of the logical terms involved?

That, in short, is the case for thinking that each of these tolerance principles should be rejected. As a consequence, any inference (other than from a sentence to the very same sentence) might be informative. But it does not follow from this that all inferences are informative. As in the case of other tolerance principles, it is likely that most instances are true, even though the universal generalization is false.

To solve paradoxes involving vagueness, it is not enough merely to reject tolerance principles. One has to explain why they seem so tempting in the first place. (And in deduction-sorites cases, it seems, the tolerance principles are especially beguiling.) In general, we might hold that tolerance principles are false but with a very high degree of truth; or that they are false but any counter-instances are unknowable, and hence unassertable; or that counter-instances shift from precisification to precisification, and so cannot determinately be recognized. Whichever explanation we use, a structure is required which preserves what Fine (1975b) calls the *penumbral connections*. In our case, if an inference is determinately informative, then any inference which includes it is also determinately informative. (And hence, if an inference is determinately uninformative, then any inference which it includes is also determinately uninformative.)

Jago's (2013b, 2014a, 2014b) models use proof rules as links between worlds, rather than as closure principles on worlds. To simplify somewhat: a proof rule directly connects a world where all the premises but not the conclusion is true to a world that's exactly the same, except that the conclusion is true too, according to that world. This connection is directed, from the 'premise' world to the

'conclusion' world. Rules with two premises connect two premise worlds to a conclusion world. (Since the worlds in question cannot be logically closed, they are all impossible worlds. But they need not be inconsistent: they could be consistent but incomplete.)

If our proof rules are taken from the sequent calculus, then there's a very direct relationship between proof rules and world-connections. To each world w we can associate two sets of sentences, $|w|^+$ and $|w|^-$: those that are true, according to w and those that are false, according to w, respectively. Then each sequent rule of the form

$$\frac{\Gamma_1 \vdash \Delta_1}{\Gamma_2 \vdash \Delta_2}$$

generates a connection from w_2 to w_1 when $|w_1|^+ = \Gamma_1 \cup \Gamma_2$, $|w_1|^- = \Delta_1 \cup \Delta_2$, $|w_2|^+ = \Gamma_2$, and $|w_2|^- = \Delta_2$. (Note how the connection goes from lower to upper sequent, for this is how, in practice, sequent proofs are constructed.) Rules with two upper sequents generate two of these connections (from the lower to each of the upper sequents).

Chaining these connections together gives a connected graph on worlds. That total graph is a *tree* (that is, a connected, acyclic graph, so that any two worlds are connected by exactly one path). Some of its *subtrees* (those parts of the whole graph that are themselves trees) correspond to proofs. That happens for a subtree T when three conditions are met:

(9.1) For each *leaf-world* w (the world found at the end of some branch) of T, $|w|^+$ and $|w|^-$ overlap (so that some A is a member of both);

(9.2) Every non-leaf node of T has at most two edges leading away from it; and

(9.3) There are edges $\langle\langle w_1, w_3 \rangle\rangle$ and $\langle\langle w_2, w_3 \rangle\rangle$ in T only if the proof system contains a rule-instance:

$$\frac{|w_1|^+ \vdash |w_1|^- \qquad |w_2|^+ \vdash |w_2|^-}{|w_3|^+ \vdash |w_3|^-}$$

These subtrees are *world proofs*. In effect, they uncover any hidden contradictions in an inconsistent and incomplete world, by connecting it to blatantly inconsistent worlds, according to which some A is both true and false. Such blatantly inconsistent worlds aren't epistemically possible for any agent, and so can't play a role in our notion of epistemic content. But other inconsistent worlds may do so, if their inconsistencies are buried deeply enough. (This raises an important worry: don't some people, rightly or wrongly, believe contradictions? We'll defer our discussion until §10.6.)

Our proposed epistemic possibility condition goes like this:

(EP) World w is epistemically possible just in case w isn't the root of any small world-proof.

This is an *absolute* notion of epistemic (im)possibility. If w is the root of any small world-proof, then it is *deeply* epistemically impossible, and not just epistemically impossible for some agent or other. Deep epistemic impossibilities are not eligible for playing a role in epistemic notions of content, and so cannot figure in our account of the content of a deduction. All the worlds not ruled out by this criterion are *deeply epistemically possible*, and together constitute *epistemic space*. (We'll discuss epistemic space in more detail in §10.3 and §10.4.)

'Small' is a vague concept. That is ultimately where vagueness enters our account of information content. If it is indeterminate whether w is the root of some small world-proof, then it is indeterminate whether w is an epistemically possible world, and hence indeterminate whether w may play a role in any content. If A is true according to w, then it will be indeterminate whether w is a member of A's content.

Let's see how this is supposed to help with the problems of logical information and informative inference. In Jago 2013b, the content of A is analysed as a pair of sets of worlds: those according to which A is true, $|A|^+$, and those according to which A is false, $|A|^-$. Call these sets the *positive* and *negative* contents of A (so that a content as a whole is a pair of a positive and a negative content). For sets Γ, we have:

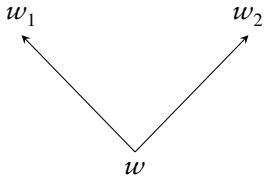
$$|\Gamma|^+ = \bigcap_{A \in \Gamma} |A|^+ \qquad |\Gamma|^- = \bigcup_{A \in \Gamma} |A|^-$$

And finally, the content of a deduction from premises Γ to conclusion A is defined as the set of all epistemically possible worlds according to which Γ is true but A is false: $|\Gamma|^+ \cap |A|^-$.

As a consequence, the clear cases of trivial inference all come out as contentless and hence uninformative. Take *modus ponens*, from $A \to B$ and A to B. Its content is defined as $|A \to B|^+ \cap |A|^+ \cap |B|^-$. Suppose this set contains a world w. Then by definition, there are sets of sentences Γ and Δ such that $|w|^+ = \Gamma \cup \{A \to B, A\}$ and $|w|^- = \Delta \cup \{B\}$. The sequent rule for $\to$ (on the left) has this instance:

$$\frac{\Gamma, A \vdash A, B, \Delta \qquad \Gamma, A, B \vdash B, \Delta}{\Gamma, A \to B, A \vdash B, \Delta}$$

and so w is the root of a world-proof:



where $|w_1|^+ = \Gamma \cup \{A\}$, $|w_1|^- = \Delta \cup \{A, B\}$, $|w_2|^+ = \Gamma \cup \{A, B\}$, and $|w_2|^- = \Delta \cup \{B\}$. This is a world-proof (and not merely a world-graph) because it is a tree, with leaf-worlds w_1 and w_2, in which $|w_1|^+$ overlaps $|w_1|^-$ (they share A) and $|w_2|^+$ overlaps $|w_2|^-$ (they share B).

This world-proof is small (by any reasonable standard of proof size). So by definition, w is not an epistemically possible world, and hence is not a member of any sentence's content. It follows that w it cannot be a member of the content of any inference, contrary to

our original assumption. So $|A \rightarrow B|^{+} \cap |A|^{+} \cap |B|^{-}$ is empty, and *modus ponens* is correctly deemed an uninformative inference by our approach. Similar reasoning applies to other clear cases of trivial inferences.

Yet not all valid deductions are deemed empty on this approach. For large n, the deduction

$$p_1, p_1 \rightarrow p_2, p_2 \rightarrow p_3, \ldots, p_n \rightarrow p_{n+1} \vdash p_{n+1}$$

is contentful. Its content consists of epistemically possible worlds according to which p_1 and each $p_i \rightarrow p_{i+1}$ $(i < n)$ are true but p_n is false. There are infinitely many worlds w for which the shortest world-proof with w at its root corresponds to $n-1$ applications of the rule for '$\rightarrow$' on the left. (To see this, just consider a world according to which nothing else is true or false, and then consider all the consistent ways of extending w.) Since by assumption n is large, all such worlds count as epistemically possible, and so the inference is contentful.

We can then formalize the notion of *trivial inference* by taking an inference to be non-trivial just in case it is contentful, in the above sense. Then a valid inference from Γ to A will be non-trivial just in case there is some epistemic possibility according to which the premises are true but the conclusion is false. So an inference is trivial just in case every epistemic possibility which represents the premises as being true does not also represent the conclusion as being false.

We will write 'triv(Γ, A)' for 'the inference from Γ to A is trivial'. To formalize the idea, we will need to be precise about which worlds count as epistemic possibilities. The simplest way to do this is to fix, artificially, precisely which world proofs are to count as the small ones. Following Jago (2014b), we pick an integer n, and say that all world proofs of size $m \leq n$ are small. Then, relative to our chosen n, we can say precisely which worlds are epistemically possible, and hence which inferences are trivial. We'll then write 'triv$_n(\Gamma, A)$' for 'relative to our chosen n, the inference from Γ to A is trivial'.

Now for the formal details. We use the standard propositional language $\mathscr{L}$ from before, with each connective $\neg$, $\wedge$, $\vee$, $\rightarrow$, and $\leftrightarrow$ as a primitive (undefined) symbol.

Definition 9.1 (Models) *A model is a tuple* $M = \langle\langle W, N, \rho\rangle\rangle$*, where* W *is a set of worlds,* $N \subseteq W$ *is the subset of normal worlds, and* ρ *is a valuation relation (as in §5.4), relating (i) each atomic sentence to exactly one truth-value at worlds in* N *and (ii) each sentence to zero, one, or two truth-values at worlds in* $W - N$*. A pointed model is a pair,* $\langle\langle M, w\rangle\rangle$ *where* w *is a world in* M*. We abbreviate* $\langle\langle M, w\rangle\rangle$ *to* M^w*.*

We then extend ρ to all sentences at worlds in N via the standard recursive clauses, as in §5.4. Then for worlds $w \in N$, $\rho_w A1$ iff not $\rho_w A0$, whereas for $w \in W - N$, ρ_w behaves arbitrarily.

Definition 9.2 (Rank) *Given a model* $M = \langle\langle W, N, \rho\rangle\rangle$ *and a world* $w \in W$*, we define* w*'s rank,* $\#w$*, as the size (number of nodes) in the smallest world-proof rooted at* w*, if there is one, and* ω *otherwise. The rank of model* M *is* $\min\{\#w \mid w \in W\}$*.*

Intuitively, a model counts as an epistemic space when its rank is not small. If we select n as our artificial precisification of 'small world proof', then only models of rank $r > n$ count as epistemic spaces.

Definition 9.3 (Trivial consequence) *For any* $n \in \mathbb{Z}^+ \cup \{\omega\}$*,* A *is an* n*-trivial consequence of* Γ*,* $\mathrm{triv}_n(\Gamma, A)$*, if and only if, for all pointed models* M^w *of rank* $r > n$*:* $\rho_w B1$ *for each* $B \in \Gamma$ *only if not-*$\rho_w A = 0$*.*

As a definition of (a kind of) consequence, this definition is rather unusual. This is because trivial consequence is not purely about truth-preservation across all epistemic scenarios. In fact, *no* inference (other than *identity*, $A \vdash A$) is preserved across all epistemic scenarios. Rather, a consequence counts as trivial in the current sense when the truth of the premises guarantees *avoidance of falsity* for the conclusion across all epistemic scenarios.

Because of this, $\mathrm{triv}_n(\Gamma, A)$ behaves as a consequence relation in some ways, but not in others, as the following results highlight. (For proofs, see Jago 2014b.)

Theorem 9.1 triv_n *has the following properties, for all* $n \geq 1 \in \mathbb{Z}^+ \cup \{\omega\}$*:*

(a) $\mathrm{triv}_n \subseteq \mathrm{triv}_{n+1}$*: if* $\mathrm{triv}_n(\Gamma, A)$*, then* $\mathrm{triv}_{n+1}(\Gamma, A)$*.*

(b) triv_n *is monotonic: if* $\mathrm{triv}_n(\Gamma, A)$ *and* $\Gamma \subseteq \Delta$ *then* $\mathrm{triv}_n(\Delta, A)$*.*

(c) $\mathrm{triv}_n(\Gamma, A)$ *only if* Γ *classically entails* A*.*

(d) triv_n *is reflexive.*

(e) $\mathrm{triv}_0(\Gamma, A)$ *if and only if* $A \in \Gamma$*.*

(f) For $n \geq 1$*,* triv_n *is non-transitive and does not satisfy cut: it is not the case that if* $\mathrm{triv}_n(\Gamma, A)$ *and* $\mathrm{triv}_n(\Gamma \cup \{A\}, B)$ *then* $\mathrm{triv}_n(\Gamma, B)$*.*

So long as n is not too small, the trivial consequences (so defined) include all the inferences we usually call trivial. Table 9.1 gives some examples, showing the minimal value of n for which triv_n holds. It is not the case that $\mathrm{triv}_3(\{A \vee B, \neg A\}, B)$ holds, for example. To see why, consider a model M containing a single world $w \in W - N$ such that $\rho_w(A \vee B)1$, $\rho_w(\neg A)1$, and $\rho_w B0$ (and that's it for ρ). Then M is of rank 4 and represents the premises as being true, but the conclusion as being false. That's all we need for the inference not to be 3-trivial.

$\mathrm{triv}_2(\{A \wedge B\}, A)$	$\mathrm{triv}_3(\{A, B\}, A \wedge B)$
$\mathrm{triv}_2(\{A\}, A \vee B)$	$\mathrm{triv}_4(\{A \vee B, \neg A\}, B)$
$\mathrm{triv}_3(\{A \to B, A\}, B)$	$\mathrm{triv}_5(\{A \to B, \neg B\}, \neg A)$
$\mathrm{triv}_7(\{\neg(A \wedge B)\}, \neg A \vee \neg B)$	$\mathrm{triv}_8(\{\neg(A \vee B)\}, \neg A \wedge \neg B)$

Table 9.1: SOME TRIVIAL CONSEQUENCES

The notion of trivial inference is interesting in its own right, but it also plays an important role in our account of fine-grained epistemic and doxastic states. We'll return to the idea in §10.5.

9.6 What Is Said

We've been investigating the notion of informative logical reasoning. This is one useful notion of content: the information contained in a valid deduction. But it clearly isn't the only useful notion of information content. Here's another: the information conveyed in a speaker's *saying that such-and-such*. This is the concept of *what is said* in that utterance. There are good reasons for thinking that this notion of content is distinct from the notion we've just been discussing.

To get a handle on what we have in mind by 'what is said', consider how:

> If someone wants to say the same today as he expressed yesterday using the word 'today', he must replace this word with 'yesterday'. … The case is the same with words like 'here' and 'there'. (Frege 1956, 296)

Here, Frege is making the point that different words can be used to say the same thing. We can 'say the same thing', but at different times, by using 'today', 'tomorrow', 'yesterday', and so on, depending on what day we speak. And by the same token, we can use the very same form of words, in different contexts, to say different things. When you utter the word 'I', you say something about yourself; but when Franz or Mark uses 'I', they say something about Franz or Mark.

This all goes to show that the concept we have in mind, *what is said*, should not be conflated with utterances (or sentence tokens), nor with sentence types. What a speaker *says* is distinct from the words she utters. We use the concept of *what is said* in the sense of *what is communicated* in making a particular utterance, as opposed to the particular way in which that content is communicated.

Saying the same thing by uttering different sentences is not a phenomenon specific to indexicals (such as 'here' and 'now') or to other context-sensitive words. Even with all contextual factors accounted for, we can still say the same thing in different ways. Just consider:

(9.4a) It's sunny and hot today.

(9.4b) It's hot and sunny today.

Suppose Anna and Bec utter these in a conversation on the same day, in the same place (and with the same conversational standards for 'sunny' and 'warm' in play). Intuitively, it seems they are saying the same thing as one another. It would be bizarre to interpret Bec as disagreeing with Anna. Try this. Imagine Bec had instead replied with, '*no*, actually it's warm and sunny today'. That response would be so bizarre, we would have to interpret her meaning using other conversational clues: perhaps as wanting to emphasize the day's warmth. Either way, Bec *says the same thing* as Anna.

This is a *purely logical* case of same-saying. Anna and Bec say the same thing because they use 'and' to connect two predicates, and (we suggest) the order in which terms flank 'and' doesn't affect what is said. We'll offer more examples like this in a moment. The immediate point here is that *same-saying* is not a phenomenon generated only by context-sensitive terms (as the initial quote from Frege may have suggested).

Note that the correct explanation is *not* that speakers of (9.4a) and (9.4b) say the same thing because those sentences are logically equivalent. This explanation would have it that all logically equivalent utterances say the same thing as one another. But this is not the case, as the following example shows:

(9.5) The Liar is both true and false.

(9.6) Claims about large cardinal numbers are neither true nor false.

These utterances do not say the same thing, even though they are (classically) equivalent. It may be that (9.5)'s speaker is a dialethist, such as Priest (1979, 1987), who diverges from classical logic in rejecting the Explosion Principle (that contradictions entail arbitrary conclusions), whereas (9.6)'s speaker is a mathematical intuitionist, such as Dummett (1978b, 1993b), who rejects Excluded Middle.

Each of these philosophical positions is completely different from the other. It is absurd to think that, in stating their different philosophical positions, they say the same thing as one another. So it is not the case that, in uttering any two classically equivalent sentences, the speakers thereby say the same thing as one another. *What is said* is a hyperintensional notion of content.

Yet, as we saw with the example pair (9.4a) and (9.4b), the 'anything goes' approach to content we mentioned in §8.4 doesn't give an appropriate analysis of same-saying. Some logical relations (including the one relating $A \wedge B$ to $B \wedge A$) preserve same-saying. Here are two further pairs in which, we think, an utterance of (a) says the same as an utterance of (b):

(9.7a) Anna or Bec will pass, and Cath will pass too.

(9.7b) Either Anna and Cath will pass, or else Bec and Cath will.

(9.8a) Either Cath doesn't like Dave or she doesn't like Ed.

(9.8b) Cath doesn't like both Dave and Ed.

These strike us as clear examples of same-saying. These same-saying pairs suggest that commutativity for 'and', distributivity of 'or' over 'and', and the De Morgan equivalences, are all operations which preserve same-saying. Replacing 'and' with 'or' in (9.4) also seems to preserve same-saying, so we may add to the list commutativity for 'or'. And associativity for both 'and' and 'or' seems obviously to preserve same-saying.

The following pair is perhaps more contentious:

(9.9a) Valeria is happy.

(9.9b) It is not the case that Valeria isn't happy.

We think an utterance of either says the same as the other. Intuitionists will disagree (at least for certain cases involving double-negations). And to be sure, there may a difference in the meaning conveyed in

English between 'Valeria is happy' and 'Valeria is not unhappy'. But the difference is a shade of meaning, not a difference in literal content; and anyway, we don't find this difference present when using 'it is not the case that', as in (9.9b). So, we say, introducing or eliminating double negations preserves what is said.

We may think of these transformations algebraically, in terms of operations $\wedge$, $\vee$, and $-$ on contents. Then we can write our principles as identities between same-saying contents: $c_1 \wedge c_2 = c_2 \wedge c_1$, $-(c_1 \wedge c_2) = -c_1 \vee -c_2$, $--c = c$, and so on. The examples suggest that we have a *De Morgan algebra*: a *bounded distributive lattice*, with top element W (the set of all worlds) and bottom element $\varnothing$, where $-$ is an involution which obeys the De Morgan laws (see, e.g., Balbes and Dwinger 1975). But we don't have a full Boolean algebra, since we don't in general have $c \vee -c = W$ or $-(c \wedge -c) = \varnothing$.

We can then use what we know of such structures to generate further predictions about same-saying. One is that both $\wedge$ and $\vee$ are idempotent: $c \wedge c = c \vee c = c$. And this seems intuitively right to us:

(9.10a) Valeria is happy.

(9.10b) Valeria is happy and Valeria is happy.

(9.10c) Valeria is happy or Valeria is happy.

all seem to say the same thing as each other (although the use of (b) and (c) would call for a rather strange context, and so may pragmatically convey more than an utterance of (a) alone).

Another consequence of this approach is that $c_1 \wedge (c_1 \vee c_2) = c_1 \vee (c_1 \wedge c_2) = c_1$, and so it predicts that each of

(9.11a) Bertie is snuffling, and either he's snuffling or Lenny is barking.

(9.11b) Either Bertie is snuffling, or he's snuffling and Lenny is barking.

(9.11c) Bertie is snuffling.

says the same as the others. But this strikes us as a poor prediction: (9.11a) and (9.11b) both say something about both Bertie and Lenny, whereas (9.11c) says something about Bertie only. So, it seems, (9.11c) can't say the same thing as the others. If that's right, then the algebra in question is weaker than a distributive lattice.

Now let's bring the discussion back to impossible worlds. If we are to understand these same-saying contents as sets of worlds, then we must restrict the worlds in question to those that obey these operations. If $A \wedge B$ is true according to one of the worlds in question, then $B \wedge A$ must be too (and so on for all the other principles just discussed). To put things another way, the logic generated by the algebra in question gives us closure conditions on worlds. Worlds which do not meet those closure conditions are not eligible for inclusion in same-saying contents. Those worlds still exist; it's just that we ignore them when we theorize about same-saying. Whatever we deem the right algebra for same-saying contents, we can model it using fine-grained worlds.

There is an objection to this approach, however. It seems that the notion of a world, in and of itself, is doing little theoretical work in this approach. All the work is done by selecting the right algebra. For given that algebra, it doesn't really matter what kind of entity the c_1s and c_2s are. Formally, they could be sets of root vegetables, and the approach would work just as well as if they were sets of worlds.

The objection then continues: surely a better approach is to select some kind of conceptual tool which *generates* the appropriate algebra, rather than being generated by it? Jago 2018b presents an alternative approach, on which same-saying contents are *sets of truthmakers*. Then two utterances say the same as each other just in case whatever would make one true would also make the other true, too. On that approach, the logic of same-saying equates to *strict truthmaker equivalence* (Fine and Jago forthcoming). This generates all the same-saying equivalences we viewed positively above, but does not imply that either $A \wedge (A \vee B)$ or $A \vee (A \wedge B)$ says the same as A.

We can always mimic this approach using impossible worlds, by focusing on only those worlds which are closed under strict truthmaker

equivalence. The objection, however, is that understanding contents in terms of truthmakers generates the right results automatically, without having to impose further restrictions (in the form of closure conditions) on worlds. In that sense, it might be said to afford us a better understanding of same-saying.

We'll consider two responses on behalf of the impossible worlds approach. The first claims that the truthmaker approach doesn't (clearly) give the best results after all. In strict truthmaker logic, $(A \vee B) \wedge (A \vee C)$ does not entail $A \vee (B \wedge C)$, and so distribution of $\vee$ over $\wedge$ is not an equivalence. Yet utterances of the following pair seem to say the same as one other:

(9.12a) Either Anna will pass, or both Bec and Cath will pass

(9.12b) Anna or Bec will pass; and also, Anna or Cath will pass.

We'll grant that it's hard to come to a clear view on this. But if we agree that they do say the same thing, it follows that strict truthmaker equivalence isn't the logic of same-saying.

The second is a metaphysical response. Contents exist, and hence we may existentially quantify over the constituents of those contents. If those constituents include possible states of affairs, then we seem to be saying that merely possible states of affairs genuinely exist. (This is the stance of the genuine realist about merely possible worlds, applied to states of affairs.) This cannot be right. If Franz is in Amsterdam, he's not in Italy. Both states of affairs are possible, but at most one exists. If both existed, then Franz would be both fully in Amsterdam and fully elsewhere. That's impossible. We shouldn't accept a semantic theory which requires reality to be like that. There are a number of responses available. One could go in for ersatz states of affairs, or non-obtaining states of affairs, or non-existent states of affairs. But all of these face issues of their own. (See Jago 2018b for an in-depth discussion of the options.)

A positive reason for analysing same-saying contents in terms of possible and impossible worlds, rather than in terms of possible states of affairs, is that it will then integrate well with other notions

of content (such as the information contents of §9.5 and doxastic contents of §10.5). All such contents are defined on the domain of worlds. This allows us to speak of one content including, overlapping, or being disjoint from another, even when they are different kinds of content (informational and same-saying, say). This is important when we want to investigate, say, how the content of what a trusted speaker says affects what their hearer thereby comes to believe.

Chapter Summary

We conceptualize information in terms of ruling out scenarios (§9.2). We discussed informative identity statements, which give rise to Frege's puzzle (§9.3), and the problem understanding how a valid logical inference can be informative (§9.4). We gave an analysis of informative logical inferences in §9.5, on which the content of a valid deduction is often indeterminate. A consequence is that it is indeterminate exactly which logical inferences are informative. We then analysed a rather different notion of content, concerning *what is said* by a speaker in making an utterance (§9.6).

10

Epistemic and Doxastic Contents

10.1 Belief States

Analyses of belief often focus primarily on individual beliefs, analysed as a relation between an agent and a sentence or proposition. But this isn't the only approach. One may, instead, focus on the agent's total belief state. That approach is appealing if we follow Dennett (1987) and Stalnaker (1984) in analysing belief as a dispositional, functional state, essentially tied to action. Recall (§8.2) Stalnaker's idea that

> Representational mental states should be understood primarily in terms of the role that they play in the characterization and explanation of action. … Our conceptions of belief are conceptions of states which explain why a rational agent does what he does. (Stalnaker 1984, 4)

An agent who is disposed to act in certain ways in a range of possible circumstances thereby has certain beliefs and desires. Being so disposed to act is precisely what it is to have those beliefs and desires, on this view. But we can't assign belief-desire pairs on the basis of single actions. There's any number of belief-desire explanations of Anna's taking an umbrella this morning. (Does she want to avoid getting wet? Does she want a makeshift weapon to hand? Does she merely like the look of carrying an umbrella, whatever the weather?)

When we factor in the agent's overall pattern of behaviour, by contrast, we can attempt a general rational explanation by ascribing

beliefs and desires. The important point here, for present purposes, is that belief is ascribed holistically, as part of a rational explanation of the agent's behaviour. So a philosophical explanation of belief should begin with holistic belief states. Individual beliefs are found further downstream.

The worlds approach model of belief (which we met in §1.2 and Chapter 5) naturally gives us an analysis of a total belief state, in the first instance. An agent's total belief state can be conceptualized, at a high level of abstraction, as a function on circumstances. The function categorizes those circumstances, depending on whether the agent can rule out a given circumstance. A state of complete ignorance corresponds to a function which, like a huge shrug of the shoulders, rules no circumstance out. Acquiring some information (or at least, what the agent takes to be information) corresponds to ruling out some of these potential circumstances. The agent's state is effectively saying, things aren't like *that*. (This way of thinking about things naturally ties in with the worlds approach to information from Chapter 9. We'll say more about the relationship in §10.5.)

This is a very general approach to thinking about representational states, and we can use it to conceptualize both knowledge and belief. Let's revisit our weather example from §9.2, in which we discussed the Bar-Hillel-Carnap theory of semantic information (Bar-Hillel and Carnap 1953). Suppose you consult the UK Met Office website and discover that it often rains in Manchester, something you previously didn't know. You then form the belief that it often rains there and, let's suppose, this belief counts as knowledge. Prior to consulting the Met Office, you had no idea about Manchester's average rainfall. As far as you were concerned, it might rain seldom there, as in Cambridge. But now you know better: Manchester is much wetter than Cambridge.

In your previous state, scenarios in which Manchester's climate is drier than Cambridge's were *epistemic possibilities* for you. If asked whether Manchester is as those scenarios say it is, you might have said 'it might be, for all I know'. This shrugging, non-committal attitude towards a scenario is what we mean by saying that the scenario is epistemically possible for you at that time. The epistemically possible

scenarios are the ones that represent a way the world might well be, for all you know.

Now that you know that it often rains in Manchester, however, those scenarios in which Manchester is mostly dry are no longer epistemically possible for you. In gaining your new piece of knowledge, you ruled out those scenarios as ways the world might be. Gaining new knowledge goes hand in hand with ruling out various scenarios as ways the world might be, as far as you're concerned.

Bar-Hillel and Carnap (1953), Chalmers (2002a, 2010), Hintikka (1962), Lewis (1975, 1986b), and Stalnaker (1976b, 1984) all develop a variation on this basic idea. It forms the basis of the Bar-Hillel-Carnap theory of semantic information (1953), on which information is conceptualized in terms of excluding certain worlds. Hintikka's (1962) semantics for epistemic logic (§5.1) can be seen in a similar light. Its crucial innovation was to add *epistemic accessibility* relations to the space of worlds. These relativize the process of ruling out worlds to specific worlds: the worlds are ruled out, from the perspective of world w, may differ from the worlds ruled out from the perspective of some other world w_1. As we put things in §5.1: the worlds accessible from w need not be those accessible from w_1.

We analyse an agent's total state of knowledge or belief, at world w, in terms of all worlds accessible from w. We can recapture individual beliefs or bits of knowledge by looking at what those worlds represent in common. Agent i believes that A (at w) when all worlds accessible from w, for i, represent that A. One very happy aspect of the worlds approach is the way the analyses of information and of belief (and knowledge) interact. In Chapter 9, we conceptualized information in terms of a partition on worlds, and we've just conceptualized belief in terms of epistemic accessibility relations on worlds.

How do these two approaches interact? Gaining information leads to new beliefs (and, in the right circumstances, to new knowledge). We can conceptualize this dynamic process in terms of an *update* on the agent's epistemic accessibility relations. The content of a learned piece of information restricts the agent's accessibility relations to the worlds contained in that content. The dynamics of *information update* and

the subsequent effect on belief, which we briefly introduced in §5.5, was studied by Baltag et al. (1998), Segerberg (1995), Van Benthem (2011), Van Ditmarsch (2005), and Van Ditmarsch et al. (2008). (Baltag and Renne (2016) give an overview of the field.)

What we want to bring out here is the deep connection between the philosophical view of belief as a dispositional, functional state, and the worlds semantics. In making the case for the semantics, we may appeal to its usefulness in the sciences as well as to its philosophical underpinning. Not many theories have it so good, and we shouldn't pass up on those benefits lightly.

10.2 The Impossible Worlds Approach

The main drawback when this approach to belief is presented in terms of possible worlds is the *logical omniscience problem*, which we discussed in some detail in §5.1. The objection hits the philosophical analysis of belief states just as hard. By now, our response should be no surprise: we suggest an analysis of belief (and knowledge) states in terms of possible and impossible worlds. That is, we maintain the analysis of belief (and knowledge) in terms of epistemically possible scenarios, but we do not require that these be metaphysically or logically possible. What seems possible, from the cognitively limited perspective of a real-world agent, need not be metaphysically or even logically possible. Epistemically accessible worlds must *seem* possible to the agent in question but, as we discussed in §9.5, an impossible world may well seem possible.

On reflection, there's nothing in the philosophical motivation for the dispositional, functional account of belief states which rules out this impossible worlds approach. On Stalnaker's view, we analyse belief states ultimately in terms of our conception of rational action, where:

> What is essential to rational action is that the agent be confronted, or conceive of himself as confronted, with a range of alternative possible outcomes of some alternative possible actions. (Stalnaker 1984, 4)

Although 'rational' plays an essential role in Stalnaker's explanation, 'possible' does not. All the work is done by 'alternative outcomes', so long as we understand the range of alternatives in question in a broad enough way. (We'll discuss whether the impossible worlds approach adequately captures the *rational* component in §10.3.)

We're going to discuss two philosophical features of the impossible worlds approach to belief states: what it says about Frege's problem, which we presented in §9.3, and what it says (or doesn't say) about scepticism.

In §9.3, we asked how it is that 'James Newell Osterberg is Iggy Pop' is informative, when 'Iggy Pop is Iggy Pop' is not. There is an analogous problem for belief and knowledge: how can one believe (or know) that Iggy Pop has turned 70, without thereby believing (or knowing) that Osterberg has turned 70? Since Osterberg is Iggy Pop, why are we not entitled to replace 'Iggy Pop' with 'Osterberg', inferring belief (or knowledge) that Osterberg has turned 70?

In general, Frege's problem of belief concerns the failure of the inference from 'x believes that Fa' and '$a = b$' to 'x believes that Fb'. The problem is to account for the failure, within a semantically plausible general theory.

The impossible worlds approach solves the problem smoothly. Suppose that a is b. Then it's necessary that a is b, and so metaphysically impossible that a and b are distinct. Given Nolan's Principle (or one of the stronger principles) from §8.4, we infer that there are impossible worlds according to which a isn't b. We may then reasonably assume that there is an impossible world w according to which a is F, but which doesn't say that b is F. (This is implied directly if we assume either (8.6) or (NP^+) from §8.4.) Then, by taking world w to be an epistemic possibility, we can model agents as believing that Fa but not that Fb. On this approach, the inference from 'x believes that Fa' and '$a = b$' to 'x believes that Fb' fails.

This solution falls out of the impossible worlds approach in a natural way. If an agent believes something under her a-concept, but not under her b-concept, then it must be acceptable both to theorize in terms of a-representations which are not b-representations, as in

the case of our world *w*, and to take that world to be epistemically accessible for that agent. For in general, we informally understand 'epistemically accessible world' in terms of what could be the case, for all the agent knows. If she doesn't know that *a* is *b*, then some world according to which *a* isn't *b* is accessible to her. The solution falls out of our general understanding of epistemic accessibility.

Now let's consider what consequences, if any, the impossible worlds approach (as applied to knowledge) has for scepticism. You know that, if you're currently reading this book, then you're not in some devilish sceptical scenario. You're not having some super-realistic dream whilst really in bed, you haven't been reduced to a brain hooked up to a hallucination-machine by an evil scientist, or anything like that. These are good things to know. But the antecedent is correspondingly hard to know. For if you know that, then you know that you're safe from scepticism.

The key premise in this kind of reasoning, which went by without much fanfare, is *Closure under Known Implication*, one version of which we already met in §5.1 as closure condition (C6):

(C6) If an agent knows that *A* and knows that *A* implies *B*, then she knows that *B*.

Those of a knowledge-positive disposition then reason by *modus ponens*. You know you're reading this book. So, you infer, you're not in a sceptical scenario. Scepticism is defeated! But this seems too easy. How, given the mere possibility of sceptical scenarios, are you so sure that you know you're in fact reading this book? The knowledge-negative alternative is to reason by *modus tollens*: since you don't know that you're not in a sceptical scenario, you don't know that you're currently reading this book. These alternatives, *dogmatism* via *modus ponens* on (C6) or *scepticism* via *modus tollens*, seem equally unattractive.

A third approach is to deny (C6). Then you seem to have the best of both worlds: knowing that you're reading this book, even though you can't in general rule out sceptical scenarios. Both Dretske (1970)

and Nozick (1981) make this move. According to Dretske, Mark knows he's been writing with Franz, even though he can't completely rule out a (wildly implausible) alternative, featuring a convincing Franz-impersonator, set on co-authoring a book about impossible worlds. According to Nozick, Franz knows where his house keys are, even though he can't completely rule out the (wildly implausible) scenario in which they've been stolen and replaced by useless replicas.

For Nozick, one's belief (and for Dretske, one's evidence) that *A* must be sensitive in the proper way to the truth of *A*. They express this idea counterfactually: had *A* been false, one would not have believed (have had evidence for) *A*. Had Mark not been writing with Franz, he wouldn't have believed he was. (That's true because the counterfactual selects the closest antecedent worlds, and these are worlds free of Franz-impersonators, in which Mark and Franz never agree to write together. Franz-impersonators belong to distant worlds.) In each case, (C6) fails. Good riddance, say Dretske and Nozick.

An impossible worlds approach to knowledge can be a hostile environment for closure principles on knowledge, including (C6). We showed in Chapter 5 how impossible worlds can be put to use to invalidate closure principles. All that's needed for a closure principle to fail is a single epistemically possible world in which the known premises hold but the putatively known conclusion does not. In particular, all that's needed for (C6) to fail is a world not closed under *modus ponens* to count as epistemically possible for some possible agent. We argued in §8.2 that there are such epistemic possibilities, and so we deny (C6). The stage seems set for us to add 'solving scepticism' to our list of benefits of the impossible worlds approach.

That would be far too quick. Epistemic closure fails, we've argued, because epistemic agents lack the cognitive resources to draw out all consequences of what they know. Take our chess players from §8.2. One player has a winning strategy available to her, yet doesn't know how best to proceed. That's because working out the winning strategy is just too complex. Compare this with a task the agent can *easily* compute. Nothing we've said so far speaks to *that* case. We

deny closure in hard cases, but this leaves open the possibility of closure for easy cases.

Spelling out 'closure for easy cases' isn't straightforward. Suppose we did it as follows:

(10.1) If an agent knows that *A* and that *A* implies *B*, and there's an easy argument from '*A*' to '*B*', then she knows that *B*.

The problem here is that easiness isn't transitive. There may be an easy argument from A_1 to A_2, from A_2 to A_3, and so on, right through to some A_n. Chaining all these arguments gives us an argument from A_1 to A_n; but this may not itself be an easy argument. But if easy arguments bring closure in their wake, then so will the long, difficult arguments we get by chaining them. Indeed, if we restrict 'argument' to 'deductively valid argument', every argument is a chain of easy arguments, and so 'easy' closure entails full closure. So this attempt to limit closure to 'easy' cases hasn't got us anywhere.

Here's a better way to capture a restricted version of closure:

(10.2) If an agent knows that *A* and that $A \rightarrow B$, then she is in a position to know that *B*. If she then competently deduces and hence comes to believe that *B* on this basis (whilst retaining what she knows), she thereby knows that *B*.

Hawthorne (2005, 29) and Williamson (2000, 117) endorse a similar epistemic closure principle. We find this weak closure principle plausible. It says that competent deduction preserves knowledge. It's hard to see how things could be otherwise. This principle, weak as it is, is strong enough for scepticism to bite. For suppose you competently deduce as above:

(10.3) I'm reading *Impossible Worlds*.

(10.4) If I'm reading *Impossible Worlds*, then I'm not in a sceptical scenario.

(10.5) Therefore, I'm not in a sceptical scenario.

Yet, it seems, you don't know that conclusion. Then you can't have known both premises. But (10.4) is a priori and easily established; it's hard to see how you could fail to know it, given that (as a rational person) you believe things to be so. Conclusion: you don't know (10.3). Scepticism prevails.

We're not arguing for scepticism. We think we know stuff. Our point is merely that the impossible worlds approach as such doesn't help with the deep philosophical issue of scepticism. Indeed, we'd be dubious of any formal semantics that claimed to do better. Scepticism is a deep philosophical problem, which requires a philosophical solution.

10.3 The Problem of Bounded Rationality

Any attempt to deal with the knowledge and beliefs of rational but non-ideal agents faces a deep problem. What must they know (or believe), given what else they know (or believe)? This is the issue of the *granularity* of epistemic and doxastic states, which we encountered in §8.4. There, we argued that there are no non-trivial closure conditions on epistemic states. So we reject closure conditions (C1–C2) and (C4–C8) for knowledge, and (C1)–(C8) for belief. ((C3) is valid for knowledge simply because knowledge is factive, which guarantees that one cannot know contradictory things.) This ensures that epistemically possible worlds include logically impossible worlds, some of which are not closed under *modus ponens*.

We also accept a weak epistemic closure principle, (10.2). But this in itself tells us nothing directly about the nature of epistemic possible worlds. It tells us that certain agents are *in a position* to know something, not that they do in fact know it. That's compatible with epistemically possible worlds obeying no purely epistemic closure principles whatever. (Rather, we should capture (10.2) in terms of a constraint on accessibility relations: doxastic accessibility must align with epistemic accessibility to the extent required by (10.2).)

The difficulty we find in accounting for rational yet non-ideal agents is that, from the theorist's point of view, the *rational* component seems in deep conflict with the *non-ideal* component. Since such agents are non-ideal reasoners, they don't know all that follows from what else they know. Yet since they are rational, the 'anything goes' approach, on which knowing something does not imply knowing anything else in particular, would seem inappropriate.

It's sometimes said that rational but non-ideal agents know whatever follows *easily* from what they know. Chalmers (2010), for example, says that

> it is plausible that p is epistemically possible when one could not *easily* come to know that $\neg p$ given what one already knows. The corresponding notion of deep epistemic possibility is something like the following: it is deeply epistemically possible that p when $\neg p$ is not easily knowable a priori. (Chalmers 2010, 105–6)

This is a natural thought. But teasing out the details is not at all simple. Suppose that one could easily come to know that A, given what one knows. Then, given Chalmers's suggestion, $\neg A$ isn't epistemically possible. If this implies that A is epistemically necessary (for that agent, given what she already knows), it immediately follows that she knows that A. This gives us the 'easy closure' principle, (10.1) from §10.2: the agent knows whatever follows easily from what she knows. Yet this can't be right, as we saw in §10.2, for easiness isn't transitive. Accepting (10.1) forces us to treat our agent as knowing whatever follows deductively from what she knows, given that any deductive reasoning can be broken down into discrete easy chunks.

Any closure principle on knowledge is vulnerable to this kind of reasoning. That's the nature of closure principles: they are all-or-nothing beasts. But 'all' is too much closure, and 'nothing' is too little rationality. This is what we will call (following Jago (2014a)) the *problem of bounded rationality*: the conflict between normative principles of rationality which govern concepts like belief, and our limited cognitive resources.

Let's take a moment to review just what rationality seems to require of belief states. Why is it, exactly, that the 'anything goes'

approach seems inadequate? Davidson (1985), Dennett (1987), Stalnaker (1984), and many others emphasize how belief ascription is a normative practice, whose purpose is ultimately to make rational sense of action. To ascribe a belief, we must first treat the agent in question as a rational being. It makes no sense to attribute, say, beliefs that A and that B, without thereby (perhaps implicitly) ascribing the belief that $A \wedge B$. Moreover, on this view, there's no question of what a belief is, outside this normative practice. As a consequence, belief simply cannot be as fine-grained as the 'anything goes' approach allows. The deep objection to the 'anything goes' approach to belief (and cognate concepts), therefore, is that the analysis fails to capture its essential rational basis.

The problem isn't limited to analysing belief. The *problem of rational knowledge* (Jago 2014b) is that the following platitudes are incompatible:

(i) Rational agents seemingly know the trivial consequences of what they know,

(ii) Rational agents do not know all logical consequences of what they know.

Here, we're using 'trivial' (as we did in in §9.5) in a way that includes all the basic inference steps (such as an instance of *modus ponens* or Disjunction Introduction), but which doesn't include all valid inferences, many of which are highly non-trivial.

The problem is that 'trivial', like 'easy', isn't transitive. Closure under trivial consequence brings full closure in its wake. So (i), interpreted as a closure principle, directly conflicts with (ii). Note that, in this formulation, the logical omniscience problem is just one half (ii) of the problem of rational knowledge. The deep problem is to avoid logical omniscience *without losing sight of the agent's rationality*.

Bjerring (2010, 2012) argues that no solution to these problems is possible. He teases out requirements that epistemic space, the space of all epistemically possible worlds, should satisfy. He then shows

that they are jointly inconsistent. We can set out the essence of his argument as follows. Take any world. If it fails some trivial inference, such as representing both A and B but not $A \wedge B$, then it is blatantly not a way the world might be. Even rational agents with very limited resources recognize this. So such worlds are not epistemic possibilities for any agent. They should play no role in epistemic space. It follows that all worlds in epistemic space, the *deeply* epistemically possible worlds, are closed under all trivial inferences. But, as we've already seen (§9.5), closure under trivial deductive consequence generates full deductive closure.

An adequate response to the problem must argue that such worlds are (deep) epistemic possibilities, even though they are not closed under trivial (or easy) inference. Our approach (as in Jago (2014a)) is to distinguish sharply between what a world represents as *not being the case*, on the one hand, and what the world does not represent as being the case, on the other. Ordinary possible worlds are maximally consistent, and so do not distinguish between not representing something as being the case and representing something as not being the case. Our worlds, however, can be incomplete. So, to represent A as not being the case, a world must explicitly say *it is not the case that* A, or $\neg A$. Contrast this with a world which is silent on A: it says neither that A, nor that $\neg A$.

Following Jago (2014a), we argue that whether a world is deeply epistemically possible depends only on what it represents. Worlds are debarred from being deep epistemic possibilities when they represent a blatant impossibility as being the case. Blatant impossibilities include blatant contradictions. So, a world which represents that $A \wedge \neg A$, or which represents both that A and that $\neg A$, is ruled out. It is not deeply epistemically possible: it cannot be treated as an epistemic possibility, for any possible agent. Similarly, if an agent can easily infer $A \wedge \neg A$ from what a world represents, then that world is ruled out. (We discussed one way to make sense of this idea in §9.5. We'll return to the idea in §10.4.)

Contrast this situation with the case in which one can easily infer some A from what a world represents, where that world doesn't

itself represent that *A*. Such worlds are not ruled out from contention as deep epistemic possibilities, we claim, for they do not explicitly deny *A*. They say nothing about *A* explicitly (even though *A* follows easily from what they do say). There is no tension between what they explicitly affirm and what they deny. (This doesn't affect our analysis of knowledge and belief as kinds of necessity, reflecting what's true according to all worlds accessible to an agent.)

To motivate the idea, consider how every story you've ever heard is *partial* in what it explicitly represents. For any story, there are some facts it remains completely silent on. (We'll say more on how fictions work in Chapter 11.) We're never told what Holmes had for breakfast the day he first met Watson. Either he had toast or he didn't. Yet the story doesn't say he had toast, and doesn't say he didn't. The story is partial. We don't, on that basis, reject the *Holmes*-stories as epistemic impossibilities. Rather, we treat them as partial representations of what we take to be determinate and complete states of affairs. That's how we should think of incomplete worlds when we're assessing whether they count as epistemic possibilities. So we should allow that they're deeply epistemically possible, just so long as they don't explicitly represent some blatant impossibility.

10.4 Bounded Rationality and Vagueness

We have to admit, we find puzzling the picture of bounded rationality we've described. We've argued (following Jago 2014a) that it's legitimate to treat incomplete worlds as epistemic possibilities. But as a result, we will have epistemic states which do not capture *trivial* logical truths, which can easily be recognized as valid. Incomplete worlds need not represent trivial logical truths, such as $A \vee \neg A$. Having accepted such worlds as epistemic possibilities, we find ourselves with agents who do not believe $A \vee \neg A$. How can we call such agents *rational*?

This is the problem of bounded rationality emerging again. An agent who does not believe all consequences of what she believes

must fail to believe some trivial consequence of what she believes. But we cannot *say* that she fails to believe that consequence, without thereby treating her as being irrational. (And perhaps, in that case, we should not describe her cognitive state in terms of belief at all.)

There is a parallel to the literature on *vagueness* here. (In what follows, we draw on Jago 2014b.) Consider the essay marking example from §9.5, in which each of 100 essays, ranging from excellent to dire, gets a different mark. It seems absurd to pronounce for sure that only those essays scoring (say) 55% or more were any good. Was the 54% 'not good' essay that much worse than the 'good' one scoring just 1% more? Yet we can't capture what's wrong with this by saying: if one essay was good, then so was the one scoring just 1% less. That principle entails that all or none of the essays were good, and that's clearly wrong.

One response to the problem is that, although there is a fact out there about precisely which essays were good, we can't pinpoint that fact with any precision (Williamson 1994). We just can't know enough about how 'good essay' is used to determine precisely which essays were the good ones. And since we shouldn't assert what we don't know (according to Williamson (1996, 2000)), we can't ever be in a position to assert things like 'only the essays scoring over 55% were any good'. (DeRose (1992), Hawthorne (2004), Stanley (2005), and Schaffer (2008) all support this *knowledge norm* for assertion.)

Jago (2014b) argues that something similar happens in the case of knowledge and belief ascriptions. A non-ideal agent will fail to know (or believe) some trivial consequence of what else she knows (or believes). This is an *epistemic oversight*: a particular case in which the agent fails to know (or believe) a particular trivial consequence of what she knows (or believes). But knowledge and belief ascriptions are part of a normative practice of explaining behaviour. We cannot be in a position to know, and hence can never assert, just which trivial consequence an agent fails to know (or believe). Epistemic oversights exist, but we can never put a finger on them.

The considerations here are very similar to those we met in §9.5. The idea pursued there was that the content of valid deductions is

indeterminate, because it may be indeterminate whether a world *w* is deeply epistemically possible. The related idea here is that states of belief and knowledge are themselves vague, because it's indeterminate which logical consequences of her beliefs an agent believes. These ideas are deeply connected. If it is indeterminate whether an inference is informative, then it may be indeterminate whether an agent who determinately believes its premises thereby believes its conclusion.

Belief (and knowledge) states are themselves vague. It's indeterminate what an agent believes (or knows), given what else she believes (or knows). It might be determinate that our agent believes that *A*, but indeterminate whether she believes that *B*, for some related *B*. Yet there must be constrains on the extent of this indeterminacy. Something like the following must hold:

(TRIV) If it's determinate that an agent knows that *A*, and the inference from '*A*' to '*B*' is trivial, then she cannot determinately fail to know that *B*.

A principle along the lines of TRIV brings with it considerable explanatory power, if we also accept either knowledge or determinate truth as the norm of assertion (Williamson 1994). Suppose that one may assert only what is determinately true. (This is implied by taking knowledge as the norm of assertion, since one cannot know what's indeterminate.) Then one can assert that agent *x* believes that *A*, but not *B*, only if it is determinate both that *x* believes that *A* and that she does not believe that *B*. But, if the inference from '*A*' to '*B*' is trivial, this is precisely the situation ruled out by TRIV. So, when '*A*' trivially entails '*B*', we can never assert that an agent believes that *A* but not that *B*. In general, we can never assert any failure of trivial closure in an agent's belief state.

This feature, we think, is what gives the misleading impression that belief states are closed under trivial consequence. As we have seen, they cannot be closed under trivial consequence. Yet we can never discern or assert any counter-instance. We mistakenly go from our inability to falsify the closure principle through counter-examples to its truth.

10.5 Belief and Trivial Inference

In this section, we'll provide formal models of belief (and knowledge), following Jago (2014b), which capture the idea that belief (and knowledge) states are themselves vague (§10.4). We'll then show how TRIV from §10.4 falls out of these models. This section is largely a technical exercise in showing how precise formal models can validate TRIV. We follow the general approach of Chapter 5, by imposing doxastic and epistemic accessibility relations on a domain of epistemic scenarios. Our domain is an *epistemic space* (§9.5), consisting of deep epistemically possible worlds. To model belief, we add a doxastic accessibility relation between worlds in the space for each agent under consideration. We can model knowledge by imposing further conditions (including reflexivity) on these relations. (Reflexivity ensures that knowledge implies truth.) And we can model both knowledge and belief by adding a doxastic and an epistemic accessbility relation for each agent, restricted so that the former implies the latter. (This ensures that knowledge implies belief.) To keep the presentation simple, we'll focus on models with doxastic accessibility relations only.

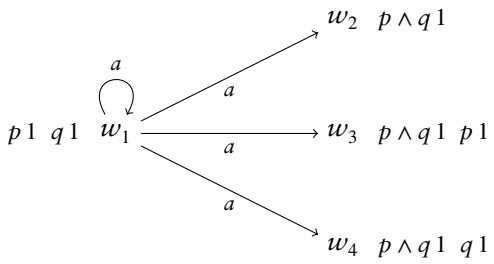
As we saw in §9.5, *epistemic space* is a vague notion. It is indeterminate just which impossible worlds count as deep epistemic possibilities, and hence indeterminate which worlds make up epistemic space. The make-up of epistemic space is governed by our analysis of deep epistemic possibility (§9.5):

(EP) World *w* is (deeply) epistemically possible just in case *w* isn't the root of any small world-proof.

A world-proof is a structure imposed by reinterpreting sequent calculus rules as relations between worlds. Its size reflects how difficult it is (in terms of number of rule-applications) to uncover a hidden contradiction within a world.

If it is indeterminate whether a world is deeply epistemically possible, then it may be indeterminate whether that world is

epistemically accessible for any agent. As a consequence, belief states may be indeterminate (just as the information contents of §9.5 may be indeterminate). Nevertheless, if we impose accessibility relations on epistemic space in the ordinary way, then many facts about an agent's belief will be determinate. As a very simple example, consider the following model, with w_1 the only possible world:



Here, w_1 is a possible world representing that p and that q, and hence that $p \wedge q$. By contrast, w_2, w_3, and w_4 are all incomplete but consistent worlds, representing just $p \wedge q$, just $p \wedge q$ and p, and just $p \wedge q$ and q, respectively.

In this model, agent a at w_1 believes that $p \wedge q$, but not that p or that q. Since both w_1 and w_2 are consistent, neither are associated with a world proof, and so, determinately, both are deep epistemic possibilities. (We argued in §10.3 that this is reasonable, for an agent may view w_2 as an incomplete description of a possible world.) So, it's determinate that our agent believes that $p \wedge q$ and determinate that she doesn't believe that p. Then there should be no problem with our asserting that she believes that $p \wedge q$ but not that p.

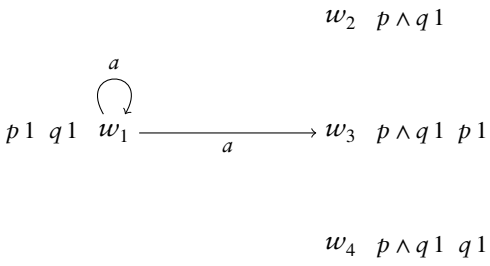
This is deeply problematic. In §10.4, we argued that there must be epistemic oversights: cases in which the agent fails to believe trivial consequences of what she believe. But we also claimed that there are no determinate cases of an epistemic oversight. They exist, but we can never put our finger on one. We are never in a position to say that our agent's belief state gives out at *this* particular point. We suggested the TRIV principle:

(Triv) If it's determinate that an agent knows that A, and the inference from 'A' to 'B' is trivial, then she cannot determinately fail to know that B.

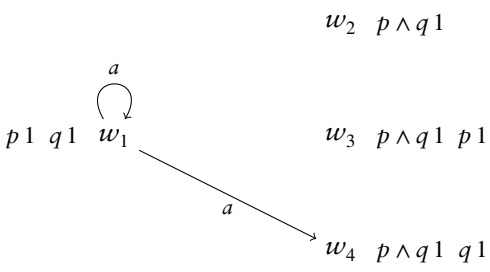
Our simple model above invalidates this principle. To capture it, epistemic models require a little more structure.

The suggestion in Jago 2014b is that an agent's epistemic accessibility relation itself is indeterminate, and the indeterminacy is not merely due to the indeterminacy of epistemic space. In our model above, all four worlds are eligible epistemic possibilities. The suggestion is that a highly incomplete word like w_2 can never be determinately accessible for an agent. A more complete version of w_2 might represent that p (as w_3 does) or that q (as w_4 does). A more complete version still would represent both that p and that q (as w_1 does).

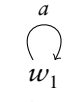
One alternative version of a's epistemic accessibility relation considers only worlds which are explicit about p, as in this model:



Another alternative considers only worlds which are explicit about q, as in this model:



What's determinate is whatever holds relative to all of these alternatives. The determinate accessibility relations, for example, are those that hold in all alternative models: from w_1 to w_1, in this case. But this doesn't mean that our model shrinks to w_1:



In this single-world model, agent *a* is logically omniscient (since w_1 is a possible world). On our approach, by contrast, it's *determinate* that some incomplete world is epistemically accessible. That guarantees that the agent's belief state isn't logically closed. But it's indeterminate which incomplete worlds are accessible, and so indeterminate where the failure of closure occurs.

(How we interpret this multiplicity of models, philosophically, is up for grabs. One option is to say that one model gets things right, but we can't ever know which. This is the *epistemicist* approach to vagueness (Williamson 1994). Another option is to say that no one model gets things right, for each model is precise in a way that reality is not. Rather, what's true amounts to what is the case according to all the models (and so equates to what's determinately true). This is the *supervaluationist* approach to vagueness (Fine 1975b). Both approaches to vagueness have their problems, of course. The aim here isn't to solve the problem of vagueness, but to bring attempted solutions to that general problem to bear on the specific problem of epistemic oversights.)

Now for the formal details, which we give for multiple agents. We use our standard propositional language $\mathscr{L}$ with knowledge operators K_i for each agent *i* and an operator '$\triangle$', read is 'it is determinately the case that …'. We define '$\triangledown A =_{\text{df}} \neg\triangle\neg A$', read as 'it is indeterminate whether …'. These operators allow our object language to express matters which are (or are not) vague. For convenience, we'll phrase our semantics for the 'K_i's in terms of epistemic projection functions (which we encountered in §5.1).

Definition 10.1 (Epistemic models) *An epistemic model for k agents is a tuple* $M = \langle\!\langle W, N, \rho, f_1, \ldots, f_k \rangle\!\rangle$, *with* W, N, *and* ρ *as in definition 9.1, and each* f_i *is the epistemic projection function for agent i, assigning a subset of* W *to each* $w \in W$. *As before, a pointed model is a pair,* $\langle\!\langle M, w \rangle\!\rangle$ *where w is a world in M. We abbreviate* $\langle\!\langle M, w \rangle\!\rangle$ *to* M^w. *The rank of* M^w *is #w, as given by definition 9.2.*

Given the precise projection functions $f_1, \ldots, f_k$, we then define the alternative projection functions, one for each sentence A for each agent i, as follows.

Definition 10.2 (A**-variant of** f_i) *Given a model M as above, we set:*

$$f_i^A w = \begin{cases} (f_i w \cap \{w \mid A \in V^+ w\}) & \text{if } f_i w \subseteq \{w \mid A \notin V^- w\} \\ \qquad\qquad \cup (f_i w \cap W^P) & \\ f_i w & \text{otherwise} \end{cases}$$

Let $f_i^{\mathscr{L}} = \{f_i\} \cup \{f_i^A \mid A \in \mathscr{L}\}$.

Definition 10.3 (Alternative sequences) *For an epistemic model M for k agents as above, let* $\mathbf{A}_M = \{\langle\!\langle g_1 \cdots g_k \rangle\!\rangle \mid g_i \in f_i^{\mathscr{L}}, i \leq k\}$. *If* $\alpha \in \mathbf{A}_M$ *is an alternative sequence, then* α^i *(for* $i \leq k$*) denotes the ith member of* α, *i.e., an alternative projection function for agent i.*

Next, we define a notion of truth for the whole language, relative to an alternative sequence. We extend the 'ρ' notation, writing '$\rho_w^\alpha A1$' to mean that A is true at w, relative to alternative sequence α.

Definition 10.4 (α**-truth**) *Given an epistemic model M as above, we define* ρ_w^α *as follows. For possible worlds* $w \in N$ *and where A is an atom, negation, conjunction, disjunction, or material implication, we set* $\rho_w^\alpha A1$ *iff* $\rho_w A1$, *with* ρ_w *as in definition 9.1. For the remaining cases (with* $w \in N$*), we set:*

(SK) $\rho_w^\alpha (K_i A)1$ *iff* $\rho_{w_1} A1$ *for all* $w_1 \in \alpha^i w$

(S△) $\rho_w^\alpha (\triangle A)1$ *iff* $\rho_w^\beta A1$ *for all* $\beta \in \mathbf{A}_M$

In all these cases, we set $\rho^{\alpha}_{w} \triangle A0$ *iff not* $\rho^{\alpha}_{w}A1$. *For impossible worlds* $w \in \mathbb{W} - N$, *we set* $\rho^{\alpha}_{w}A1$ *iff* $\rho_{w}A1$ *and* $\rho^{\alpha}_{w}A0$.

Definition 10.5 (***n*-entailment**) *Given a pointed epistemic model* M^{w} *where* $M = \langle\!\langle W, N, \rho, f_1, \ldots, f_k \rangle\!\rangle$, *A is true in* M^{w} *iff* $\rho_{w}^{\langle\!\langle f_1 \cdots f_k \rangle\!\rangle}A1$. *Then, for any* $n \in \mathbb{Z}^{+} \cup \{\omega\}$, *logical n-entailment,* $\models_{n}$, *is defined as:*

> $\Gamma \models_{n} A$ *iff, in every pointed model* M^{w} *of rank* $r \geq n$ *where* $w \in N$, *each* $B \in \Gamma$ *is true in* M^{w} *only if A is true in* M^{w}.

It is then easy to see that $\models_{n}$ extends classical (propositional) entailment:

Theorem 10.1 *For any* $n \in \mathbb{Z}^{+} \cup \{\omega\}$: *if* Γ *classically entails A, then* $\Gamma \models_{n} A$.

Proof: By contraposition. If $\Gamma \not\models^{n} A$, then for some pointed model $\langle\!\langle W, N, \rho, f_1, \ldots, f_k \rangle\!\rangle^{w}$ of rank $r \geq n$ with $w \in N$, each $B \in \Gamma$ is true but A is not. Set $VA = 1$ iff $\rho_{w}^{\langle\!\langle f_1 \cdots f_k \rangle\!\rangle}A1$. Then it is easy to see that V is a classical valuation function over atoms $\{p, \triangle A, K_i A \mid p, A \in \mathscr{L}\}$. Since $VB = 1$ for each $B \in \Gamma$ but $VA = 0$, it follows that Γ does not classically entail A. ■

Now recall the formal notion of *n-trivial consequence*, $\mathrm{triv}_{n}(\Gamma, A)$, from §9.5 (definition 9.3). Both of our key formal concepts, *n*-entailment and *n*-trivial inference, are parameterized by an integer (or ω) *n*, which we think of as an artificial precisification of vague epistemic space. Given any such precisification, we can show that our models imply TRIV: if it's determinate that an agent knows each of the premises Γ of a trivial inference from Γ to A, then it's not determinate that she fails to know that A.

Theorem 10.2 *For any* $n \in \mathbb{Z}^{+} \cup \{\omega\}$, *if* $\mathrm{triv}_{n}(\Gamma, A)$ *then* $\{\triangle K_i B \mid B \in \Gamma\} \models_{n} \neg\triangle\neg K_i A$.

Corollary 10.3 *For any* $n \in \mathbb{Z}^{+} \cup \{\omega\}$, *if* $\mathrm{triv}_{n}(\Gamma, A)$ *then* $\{\triangle K_i B \mid B \in \Gamma\} \cup \{\neg K_i A\} \models_{n} \triangledown K_i A$.

Corollary 10.4 *For any* $n \in \mathbb{Z}^+ \cup \{\omega\}$, *if* $n \geq 3$, *then* $\vDash_n \neg\triangle\neg K_i(A \vee \neg A)$ *and* $\neg K_i(A \vee \neg A) \vDash_n \triangledown K_i(A \vee \neg A)$.

Proof sketch: Suppose that $\text{triv}_n(\Gamma, A)$. Then for any pointed epistemic model of rank $r > n$ where Γ is true, A is not false. Suppose also that $\triangle K_i B$ for each $B \in \Gamma$ is true in some M^w. Then, for each alternative accessibility relation and each i-accessible world w', B is true at that world. Then A is not false at w'. But, since the alternative projection function f_i^B forces a stance on A, A is true at each world $w' \in f_i^B w$. Thus $K_i A$ is true at w relative to any alternative sequence containing f_i^B. Then since $\neg K_i A$ is false at w relative to some alternative sequence, $\triangle\neg K_i A$ is false, hence $\neg\triangle\neg K_i A$ is true, in M^w. Corollary 10.3 follows given $\triangledown A =_{\text{df}} \neg\triangle A \wedge \neg\triangle\neg A$, and corollary 10.4 is a special case. Full proofs can be found in Jago 2014b. ■

However we choose a precise delineation of 'epistemic scenario' and 'trivial consequence', if the inference from Γ to 'A' is trivial then determinate knowledge of Γ entails the agent does not determinately fail to know that A, just as our principle TRIV from §10.4 says. Equivalently (as corollary 10.3 says), if agent i does not know some trivial consequence 'A' of what she knows, then it is indeterminate whether she knows that 'A'. So, on the account proposed, there are no determinate epistemic oversights. Each case of an epistemic oversight is an indeterminate case.

Since what is indeterminate is not rationally assertible, it is then never rational to assert that agent i suffers from a particular epistemic oversight. If an agent is not logically omniscient, then we can be sure that she suffers from some epistemic oversight. Indeed, it is determinate that real-world agents are not logically omniscient, and hence determinate that real-world agents suffer from epistemic oversights. But we can never say what they are: we cannot locate them in a rational agent's epistemic state. Whenever we focus on a particular trivial consequence 'A' of agent i's knowledge, it is never rational to assert that she does not know that A (even if that's the case). In this way, our formal models support our philosophical contention

that epistemic oversights are always elusive, just as counterexamples to tolerance principles for vague predicates are.

10.6 Believing Contradictions

There's a worry that's been building up since §9.5, where we introduced our concept of *epistemic space*. There, we excluded explicitly contradictory worlds (according to which some *A* is both true and false) from epistemic space. We also excluded worlds from which we can reach explicitly contradictory worlds in a relatively short number of steps. All these excluded worlds are associated with a small world-proof, and are excluded on that basis. Our key principle is this:

(EP) World *w* is epistemically possible just in case *w* isn't the root of any small world-proof.

For some worlds, it's indeterminate whether they satisfy the criterion (because it's indeterminate whether any associated world-proof is small). But it's determinate that explicitly contradictory worlds are not epistemic possibilities (for any agent). So they played no role in our analysis of epistemic and doxastic states (§10.5).

The objection is simple: surely there *are* agents who believe explicit contradictions. Whether they're right or wrong to do so isn't the point, for beliefs needn't be true, or even reasonable. People can believe all sorts of things. There's even a well worked out philosophical view, *dialethism*, according to which contradictions can be true. And yet, we're saying that it's impossible to believe a contradiction. How can we square this?

Take the case of Graham Priest, dialethism's foremost proponent (Priest 1979, 1987, 2014, 2016b). Few people on earth are more rational and logically adept. Fewer still know more about negation. So when Priest says, clearly and repeatedly, that he believes contradictions, and backs this up with sophisticated philosophical and logical argument, how can we disagree with him?

Dialethists clearly believe something when they assert a contradiction. We follow Jago (2014a, §7.5) in thinking that their assertions mean something a little different from what they would seem to mean. In particular, the dialethist may mean something a little different from what other English speakers typically mean by 'not'. The evidence for this comes out in the logical rules she takes to govern her concept of negation. In classical and intuitionistic logic, the logical rules for 'not' allow us to link assertions to denials (or what we accept to what we reject).

Suppose in a conversation we keep track of the things we accept, Γ, and the things we reject, Δ, by writing: $\Gamma \vdash \Delta$. Then, if we accept A, we should reject $\neg A$, and vice versa:

$$\frac{\Gamma, A \vdash \Delta}{\Gamma \vdash \neg A, \Delta} \qquad \frac{\Gamma \vdash A, \Delta}{\Gamma, \neg A \vdash \Delta}$$

These are the sequent rules for classical negation. The dialethist rejects them both. Her concept of negation differs from the classical one. (She readily agrees, for she finds the classical notion incoherent.) The question, then, is what is typically expressed in English by 'not'.

The line in Jago 2014a, §7.5, then goes as follows. In communicating with each other, we need some way to indicate what we accept and what we reject (or deny). We could have evolved green and red lights on our heads, so that uttering '*A*' with a green light amounts to accepting that *A*, whereas uttering '*A*' with a red light amounts to denying it. Or we could put a thumb up or down as we utter '*A*'. But there's a simpler way: we use 'no', 'not', and cognates, to signal disagreement and denial. (That's a very useful tool, since we can then deny a denial, by using 'not' twice.)

This argument is part conceptual, part empirical. The conceptual part is that, to engage in the kind of discursive practices our complex interactions require of us, we need a mechanism for signalling acceptance and rejection of contents. The empirical part is that English speakers typically use 'not' and 'no' for that purpose. But we're free to step out of that practice. Dialethist uses of 'not' may

well express their non-classical, non-explosive concept of negation. They believe those contents, which they express using 'not' and we express using 'dialethist negation'. But those contents are not literally contradictions, because contradictions are those involving *negation*, the concept expressed in standard English by 'not'. Dialethists believe what they say when they say, for example, 'the Liar is both true and not true'. Their assertions are contentful and meaningful. But they don't believe that the Liar is both true and not true.

Chapter Summary

We outlined the case for making belief states the primary focus of our analysis (§10.1), and for including impossible (as well as possible) worlds in that analysis (§10.2). This allows us to deny various closure principles, although this probably won't help defeat worries about external-world scepticism (§10.2).

The issue that concerned us most is the *problem of bounded rationality* (§10.3): belief states seem to be closed under 'easy' trivial consequence, but not under full logical consequence, and yet the former implies the latter. Our solution was that some trivial closure principle must fail on a given belief state, yet it is indeterminate just where this occurs (§10.4). We cannot know, or be in a position that entitles us to assert, which trivial consequence of her beliefs an agent fails to believe. We gave formal models of belief states alone these lines, and showed that they respect the indeterminacy-of-closure intuition, in §10.5. Finally, in §10.6, we discussed how we might square this approach, which says that no one can genuinely believe a contradiction, with the fact that some people seem to believe just that.

11

Fiction and Fictional Objects

Co-authored with Christopher Badura

11.1 Problems of Fiction

Our ability to tell fictional stories, to engage with them, to think about their characters, and to reason about the situations they find themselves in, is an important part of being human. Tales, novels, plays, and operas represent things as being such-and-so. Besides enjoying their fictional contents, we can learn a lot from them. But how can this be, given that, in general, they are no *true* representations of reality?

In this chapter, we will set out some philosophical issues of fiction and discuss to what extent impossible worlds help with them. We will tackle two problems in particular. These can be mapped to a prima facie intuitive distinction between two kinds of fiction-related discourse. Some features are taken as true of fictional characters like Heathcliff, Gandalf, and Sherlock Holmes, *within* the fictions in which they appear. It is true in the respective fictions that Sherlock Holmes is a detective and that Gandalf is a wizard. Call this *intra*-fictional discourse.

We also ascribe to those characters features they don't have in their respective fictions. We say that Heathcliff is a fictional character due to Emily Brontë and that Holmes is more famous than any real detective. These claims are not true in the stories: within the respective

fictions, Holmes avoids celebrity and Heathcliff is not fictional at all, but a very real, tormented hero. Call this *extra*-fictional discourse.

The distinction between intra-fictional and extra-fictional discourse may not be sharp in all cases (Pelletier 2003), but it seems robust and intuitive enough. Fictional characters enjoy a double life. In Kit Fine's words:

> On the one hand, they have certain properties within the contexts in which they appear; they love and hate, thrive and fail, and live their varied lives. On the other hand, they also relate to the real world; they are created by authors, read by readers, and compared, for better or worse, with one another and with what is real. (Fine 1982, 97)

We start by focusing on one side of this double life, the intra-fictional one. What does it mean that something is *true in a fiction*? Fictions are not, on the whole, true, nor do their authors generally aim for them to be. (Recall the proviso, 'any resemblance to real people or events is purely coincidental'.) But in intra-fictional discourse we can truthfully talk about what happens in a story. We can do so even when such truths are not explicitly stated in the fiction. When Heathcliff and Catherine meet for the final time, Heathcliff is dressed in the manner of an eighteenth-century country gentleman and not as a circus clown. That's an intra-fictional truth about the world of *Wuthering Heights*. Yet the text of that scene never explicitly says anything about how Heathcliff is dressed. So what is true within the fiction of *Wuthering Heights* must go beyond what's explicitly written in the text. The problem is to explain how this is so: how it is true in *Wuthering Heights* that Heathcliff was dressed in the manner of an eighteenth-century country gentleman, and not as a circus clown, at his final meeting with Catherine.

The second problem we will discuss is that of fictional entities such as Holmes, Gandalf, Catherine, and Heathcliff. As fictional characters, it seems that they exist in the fictions of *Sherlock Holmes*, *The Lord of the Rings*, and *Wuthering Heights*, and not in reality. If this means that there really are no such things as Holmes, Gandalf, Catherine, and Heathcliff, how can we make true claims about them

in extra-fictional discourse? How can we (let's assume, truthfully) say that Holmes is more famous than any real detective, that Emily Brontë created Heathcliff, and that Kate Bush's Heathcliff is the same Heathcliff as Emily Brontë's? How can it be true that Kate Bush and Emily Brontë are speaking about the same Heathcliff, if there is no such thing?

Let us start with the first issue: truth in fiction. We have a range of intuitions about what's true in a given fiction (Woodward 2011). The issue is how to systematize them.

11.2 Truth in Fiction

We commonly talk of the worlds of fiction ('in *Star Trek*'s world …', 'in the world of *The Lord of the Rings* …'). So we might understand what's true in a fiction along the lines of truth relative to a world. But in general, many worlds (including many possible worlds) will be compatible with the explicit text of *Wuthering Heights*. Which of these worlds is the world of *Wuthering Heights*?

It's not even clear that the world of the fiction must be compatible with everything stated in the text. Sophisticated 'unreliable narrators' explicitly state something that turns out later on not to be true in the fiction. The narrator may make ironic or tongue-in-cheek remarks. Further subtleties involve the author-narrator distinction and the positing of a fictional or implicit author (Currie 1990). But since our aim here isn't to attempt a full theory of truth in fiction, we're going to ignore these subtleties. We want to focus on the role impossible worlds may play in a good theory. We shall only take into account reliable and literal narration, in which a narrator's explicitly uttering that *A* is sufficient for *A* to be true in the fiction. This seems to be the default case.

(Heyd (2006) proposes to account for unreliable narration pragmatically, via Gricean maxims. Perhaps non-literal speech and genre conventions, such as characters speaking in verse (Walton 1990), or

mentions of '555' numbers in American television (Hanley 2004), should also be dealt with pragmatically.)

The set of sentences explicitly and literally uttered by the narrator gives the *explicit content* of a fiction. We thus endorse the following principle as a default rule:

(EXPLICIT) If *A* occurs explicitly in the story *f*, then *A* is true in *f* (and so 'in *f*, *A*' is true).

The worlds complying with the explicit content of *Wuthering Heights* are all worlds in which Heathcliff is adopted by Mr Earnshaw, comes to own Thrushcross Grange and Wuthering Heights, marries Isabella, and so on. Although the explicit content of the novel narrows down the class of worlds to those compatible with it, we cannot get to a unique world we can call *the* world of *Wuthering Heights*. Fictions are *incomplete*. Given a fiction *f*, there are sentences *A* such that neither it nor its negation is true in *f*.

In an extremely influential paper, Lewis (1978) proposes that we associate a plurality of possible worlds with the fiction and then analyse truth in the fiction as whatever is true according to all those worlds. He considers three accounts of how these worlds are selected. In the first, truth in fiction is a matter of what's true at all those worlds 'where the fiction is told, but as known fact rather than fiction' (Lewis 1978, 40):

(FICTION$_1$) 'In fiction *f*, *A*' is true iff *A* is true at every world where *f* is told as known fact rather than fiction. (Lewis 1978, 41)

This approach generates too few truths-in-fiction. To understand *Wuthering Heights*, we need to understand something of the social customs of late eighteenth-century England. Heathcliff leaves, becomes wealthy, and returns as a gentleman. To understand the importance of those events, we need to understand something of England's attitudes to class and gender relations in the eighteenth century. Emily Brontë took for granted that her contemporary readers

would easily grasp these, so didn't include any explicit facts about class and gender relations in her text. But FICTION$_1$ ignores these.

There will be worlds where *Wuthering Heights* is told as known fact but eighteenth-century English social relations are turned on their heads. So, according to FICTION$_1$, it won't come out true in *Wuthering Heights* that Heathcliff exerts social power over Catherine. And there are plenty more contingent actual-world facts which should appear in the world of *Wuthering Heights*, but which are ignored by FICTION$_1$: that there's gravity, that it obeys an inverse-square law, and so on.

FICTION$_1$ generates too few truths because it includes too many worlds. Perhaps we should consider just those worlds closest to our own which are compatible with the text. This approach ignores all those which differ radically from our own, in ways not required by *Wuthering Heights*'s being told as known fact. This is Lewis's second analysis:

(FICTION$_2$) 'In fiction *f*, *A*' is true iff some *A*-world where *f* is told as known fact differs less from the actual world than any non-*A*-world where *f* is told as known fact. (Lewis 1978, 42)

This approach seems to include too much of the actual world in the world of the fiction (Currie 1990, Proudfoot 2006). In the closest worlds to ours compatible with *Wuthering Heights* being told as known fact, Jeremy Corbyn is Labour leader in 2017. So FICTION$_2$ treats the fact of Corbyn's leadership in 2017 as a truth of *Wuthering Heights*. That's rather surprising. *Wuthering Heights* is about Catherine and Heathcliff's story, not Corbyn's. The approach also fixes too much of our physical law. Since there are in fact no ghosts, FICTION$_2$ implies that any ghost story should be understood as a story in which people merely hallucinate or imagine ghosts.

FICTION$_2$ delivers strange consequences for non-realistic fiction. What kind of world is closest to our own, whilst realizing everything explicitly described in *The Lord of the Rings*? Perhaps one in which the goings on in Middle Earth take place in a bubble universe, causally isolated from our own reality, and with its own laws of nature. We

might picture a world like that as a perfect duplicate of our own reality, but with a Middle Earth-universe tacked on the side. For any such world, the smaller its Middle Earth bubble is in time and space, the more similar it is to our own reality. The closest such world is probably one in which the Middle Earth bubble vanishes from existence the moment the story closes. Yet it doesn't seem right to say that it's true in the fiction that Middle Earth annihilates an instant after Sam comes home, draws a deep breath and says, 'Well, I'm back'.

A compromise involves similarity, not with actual fact, but with the 'generally prevalent beliefs … of the author and his intended audience' (Lewis 1978, 44). (Lewis speaks of 'overt beliefs': these are what almost everyone believes, almost everyone believes that almost everyone believes, and so on.) We consider a world which realizes the overt beliefs of Emily Brontë and her readers. We then move to the closest worlds at which *Wuthering Heights* is told as known fact. Whatever is the case at those worlds is what's true in *Wuthering Heights*. This is Lewis's third analysis:

(FICTION$_3$) 'In fiction f, A' is true iff for each collective belief world w of the community of origin of f some A-world where f is told as known fact differs less from w, than does any non-A-world where f is told as known fact. (Lewis 1978, 45)

A worry here is how we balance the author's beliefs with her audience's in arriving at a 'collective belief' world (Bonomi and Zucchi 2003). Take a novel written in Nazi Germany by a progressive author, opposing the regime. Is the collective belief world one of progressive anti-Nazi feeling? Or are the author's beliefs outweighed by her audience's predominantly Nazi ideology?

We find FICTION$_3$ the most promising of Lewis's proposals. In the next section, we'll refine it using impossible worlds, together with ideas from *belief revision* theory.

11.3 Hyperintensional Fictions

In each of the analyses we've looked at so far, truth-in-fiction relies on possible worlds where the story is told as known fact. But an inconsistent (or otherwise impossible) story can't be known, and so can't be told as known fact. So what should we say about truth in inconsistent fictions?

There are different sources of inconsistent fiction. One is narrative oversight, in which the inconsistency is accidental to the plot. Watson's war wound is in his shoulder in some stories; in others, it's in his knee. It's also a given that he has only one wound. We might take the fictional truth to be given by whatever's true in some maximally consistent fragment of the story, or we might take it to be whatever's true in all of them (Lewis 1978, 46). On the former, Watson's wound is in his shoulder, and Watson's wound is in his knee, but it's not both. Truth-in-fiction is non-adjunctive: A and B can each be true in the fiction without $A \wedge B$ being true in the fiction. (This is a version of the subvaluational 'fragments of belief' approach from §8.2.) On the latter approach, it's true in the fiction that Watson's war wound is either in his shoulder or his knee, but neither disjunct is true in the fiction.

Not all inconsistent fictions are due to narrative oversight. Some are intentional; some blatantly so:

> Carefully, I broke the tape and removed the lid. The sunlight streamed through the window into the box, illuminating its contents, or lack of them. For some moments I could do nothing but gaze, mouth agape. At first, I thought that it must be a trick of the light, but more careful inspection certified that it was no illusion. The box was absolutely empty, but also had something in it. Fixed to its base was a small figurine, carved of wood, Chinese influence, Southeast Asian maybe.
>
> I put the lid back on the box and sat down hard on the armchair, my mental states in some disarray. I focused on the room. It appeared normal. My senses seemed to be functioning properly. I focused on myself. I appeared normal. No signs of incipient insanity. Maybe, I thought, it was some Asian conjuring trick.

> Gently, I reopened the box and gazed inside. … The box was really empty and occupied at the same time. The sense of touch confirmed this. (Priest 1997b, 575–6)

This is the central passage of Graham Priest's short story, *Sylvan's Box*. The narrator is Priest himself (or, a fictional version of him). Qua author, he asks us: what's true in this fiction? The most straightforward reading takes the narrator's statements at face value. It's true, in the fiction of *Sylvan's Box*, that Priest discovers a box that is both empty and not empty. The obtaining of a contradiction is essential for understanding the story. But no operation on maximally consistent fragments of the story will deliver the result that, in *Sylvan's Box*, a box is both full and empty at the same time.

Hanley (2004) and Nolan (2007) challenge the view that the contradiction is essential to *Sylvan's Box*. Both opt for a reading under which Priest, as narrator, falsely believes that there is a simultaneously full-and-empty box. His actions are understandable, in light of his false beliefs. That's surely one possible interpretation and, as such, there's nothing wrong with it. One might read Priest's text that way if taken as an attempted statement of historical fact. But we don't see why we *must* understand the story in this way. Is an author powerless to make her characters perceive and judge veridically and her narrator speak accurately? Surely not. We think the 'Principle of Poetic Licence' (Hanley 2004, 121, attributed to Harry Deutsch) is plausible:

(PPL) For any *A*, one can write a story in which *A* is true.

Other impossible fictions suggest that the inconsistent (or otherwise impossible) reading must be available. Proudfoot (2006) discusses the comedy film *Last Action Hero*, in which the central plot device involves a teenage boy, Danny, who finds himself in the fictional world of his favourite film hero, Jack Slater. Slater's evil nemesis later escapes to the real world and Danny and Jack have to follow. Throughout the film, various human characters become fictional, and various fictional characters become human. We take it that this is metaphysically impossible: each real human is essentially human and

so necessarily non-fictional. But it doesn't make sense of the plot to claim that Danny is hallucinating. If the impossible reading is available here, it should be available also for *Sylvan's Box*.

Hanley (2004) challenges our assumption, EXPLICIT, that the utterances which constitute the story are all true in that story (§11.2). (We assumed this with the proviso that unreliable narration and non-literal speech should be treated pragmatically.) But whilst EXPLICIT implies that truth-in-*Sylvan's Box* is inconsistent, we don't need to rely on EXPLICIT to make that case. Even if EXPLICIT failed for each particular contradiction occurring explicitly in a story, this would not stop implicit contradictions being true in the story. There are inconsistent time travel stories in which no contradiction is presented explicitly, for example. We may need a good deal of reasoning to unpack the inconsistency. Rather, our argument above was based on intuitive judgements about what's true in *Sylvan's Box* and in *Last Action Hero*, which are then systematized into a cohesive (but not consistent!) narrative.

Let's return to Priest's story and the morals he draws from it. It is not true in *Sylvan's Box* that Priest finds himself levitating, dressed in a tutu. Yet that would be classically entailed by a contradiction, such as the box's being empty and not empty at the same time. So what's true in this fiction isn't closed under classical entailment:

> in understanding the story one has to draw inferences – often non-monotonic ones – from what is explicitly presented, together with background information. … Clearly, however, the deductive canons employed cannot be those of classical logic. … The logic of the story must be paraconsistent. (Priest 1997b, 580)

If the logic of this fiction is paraconsistent, not supporting the inference from a contradiction to arbitrary conclusions, then the worlds of the fiction must include impossible worlds. But exactly which logic is the 'logic of *Sylvan's Box*'? It seems that *some* inferential principles must apply: *Sylvan's Box* is (explicitly) set near Canberra and so we may infer that it is set in Australia, although this is not stated explicitly. Assumed background knowledge (that Canberra is

in Australia) allows us to infer an implicit truth of the fiction from an explicit one. We infer this conclusion using the transitivity of 'located within': if x is located within y and y within z, then x is located in z. *Modus ponens* is required to make sense of that reasoning.

Other logical principles seem equally necessary to make sense of the story. In the story, the box is empty. It also has something in it. Those truths of the fiction are, strictly speaking, not explicit in the text. What's explicit is their conjunction (expressed with 'but'). Yet it seems clear that the conjuncts are also true in the fiction. Further exploration of the text may provide evidence that the standard introduction and elimination rules for conjunction, disjunction, and implication, plus double negation introduction and elimination, are all fine in *Sylvan's Box*. This might give reason to think that the logic of this fiction is a paraconsistent logic in which '$\rightarrow$' is not the material conditional. One of the relevant logics of Chapter 6 may then fit the bill.

That may be the right logic for *Sylvan's Box*. But what, in general, is the right logic of the 'in fiction f' operators? It is this logic, if one exists, which deserves the label 'the logic of fiction'. We are not sure that one logic may be singled out as *the* logic of fiction. Recall the Principle of Poetic Licence (PPL) from above: for any A, we can always create a fiction in which A is true. So, for any candidate 'logic of fiction', it seems we can create a fiction which breaks the rules of that logic (Proudfoot (2018) makes a convincing case for this).

(Actually, PPL suggests, but does not imply, that conclusion. For that, we'd need a stronger principle: that for any $A_1, \ldots, A_n$ and C, we can create a fiction in which each A_i is true but C is not. This echoes the discussion of *Nolan's Principle* (NP), and the stronger (NP$^+$), from §8.4.)

11.4 A Formal Semantics

We'll now set out an account of truth in fiction, as developed by Badura and Berto (2018). It's based on Lewis's third analysis (FICTION$_3$ from §11.2), expanded to include impossible worlds and

considerations from cognitive science (Nichols and Stich 2003), which inspired our treatment of the logic of imagination in Chapter 7. The key idea is that fiction involves a kind of *belief revision*.

The semantics for doxastic and epistemic logics of belief revision involves worlds ordered (usually, totally) by some *plausibility* relation (Grove 1988, Segerberg 1995, 2001). We can think of the ordering of worlds in terms of spheres, nested around a core, as in the standard Lewis semantics for counterfactuals (Lewis 1973b). But here, the spheres do not model objective similarity. Instead, they model subjective plausibility, or degrees of belief entrenchment. The innermost sphere is the set of worlds that realize certain beliefs. The closer other worlds are to the core, the more plausible they are for the relevant agent(s). These worlds are more likely to be embraced as fallback belief-positions, after new information induces an agent to revise her beliefs.

Roughly, we take 'in fiction f, A' to be true when, having revised the relevant agents' beliefs with f's explicit content, all the most plausible worlds are A-worlds. To achieve this, we think of revisions as *soft upgrades* (Baltag and Smets 2011, Van Benthem and Liu 2007), in which the fiction's explicit content leads to a reordering of the set of worlds. Our motivation for this comes from Currie (1990), Lewis (1978), Walton (1990), and others, who claim that, in engaging with fiction, we play a game of *pretence* or *make-believe*. (Matravers (2014) challenges that this is the crucial distinction between fiction and non-fiction. However, what matters for our purposes is just that it be one feature of fiction.)

Nichols and Stich (2003) give a cognitive model of pretence that resembles ideas from belief revision theory. Our cognitive architecture comprises a 'world box', accessed when we engage in pretence via the pretence premises, and a 'belief box', from which we pick background beliefs to integrate into our pretence. We then consider those worlds where the pretence premises obtain and that are most plausible in terms of our background beliefs, adjusted to make room for the explicit pretence premise itself. The content of the pretence is whatever holds in all such worlds.

In the preference-based belief revision semantics, agents order the set of worlds based on how strongly they take the various worlds as candidates for actuality. Let '$v \leq_w u$' stand for 'v is at least as plausible as u with respect to w' (for a given agent). In one kind of soft upgrade, new information that A reorders the worlds so that A-worlds become the minimal elements in the new ordering, $\leq_w^A$. Then, if A is true in v but not in u, $v \leq_w^A u$.

Think of an agent reading a fiction as being faced with new information, sequentially upgrading beliefs with each sentence the agent reads. One needn't actually believe what one reads, of course. Nichols and Stich (2003) suggest a mechanism for quarantining the update, by indexing worlds by their previous position in the ordering. When the pretence ends, one will recover the initial ordering from the index.

To make sense of truth in fiction, we have to consider the beliefs of a community of agents. For Lewis's FICTION$_3$, these are the overt beliefs of the community of origin of the fiction. We depart from Lewis in two ways. First, we consider the beliefs of the community of *interpretation*, rather than the community of origin. We want to model the interpreting community's reasoning on the fiction. Second, rather than overt beliefs, we take common ones: that is, the beliefs such that everyone has them, everyone believes that everyone has them, and so on. This is a merely practical choice: it replaces the 'most' quantifier over beliefs with 'all'. When we evaluate 'in fiction f, A', we look at all worlds (possible or not) that are minimal with respect to the common belief worlds of the community of interpretation after upgrading with the explicit content of f.

Now let's present the formal semantics. (For ease of exposition, we will not take into account Nichols and Stich's (2003) considerations about quarantining.) As before, we use a propositional language $\mathscr{L}$ built in the usual way using $\neg$, $\wedge$, $\vee$, modalities $\Box$ and $\Diamond$, the strict conditional $\dashv\exists$, and, for each fiction f, an 'in the fiction' operator 'In f'.

A *multi-agent plausibility model* for a finite set of agents Ag is a tuple $\mathscr{M} = \langle\!\langle W, N, \{ {}_i\leq_w \mid i \in Ag, w \in W\}, \rho\rangle\!\rangle$, where W is a set of worlds, $N \subseteq W$ is the set of normal worlds, each ${}_i\leq_w$ is agent

i's plausibility ordering on W with respect to $w \in W$, and ρ is a valuation relation (see §5.4 and §7.2), relating formulas to 1, 0, both, or neither at the various worlds. We assume each ${}_i\leq_w$ to be transitive and well-founded, so that we can always determine the most plausible worlds, $\min({}_i\leq_w, S) = \{v \in S \mid (\forall u \in S) v\ {}_i\leq_w u\}$, for any $S \subseteq W$. Well-foundedness implies reflexivity and total ordering, so that any two worlds are comparable with respect to each ${}_i\leq_w$.

Given individual orderings, it is a non-trivial task to come up with a group ordering ${}_G\leq_w$ for some group of agents $G \subseteq Ag$. Such a group ordering reflects whether the agents can agree for every set of worlds on some set of most plausible (or preferred) worlds. This amounts to a voting problem among infinitely many (or at least arbitrarily finitely many) alternatives. Our approach is therefore constrained by Arrow's (1950) impossibility theorem for social choice functions (and Leitgeb and Segerberg's (2007) analogue for belief revision).

These results show that, given certain conditions on the group of agents, a group preference ordering cannot be obtained. The reasoning assumes that any preference ordering is permissible. In our case, however, an upgrade requires that the worlds where the story is told as known fact be considered at least as plausible as any world where the story is not told as known fact. So Arrow's impossibility result does not straightforwardly apply in our case. We will assume a transitive and well-founded group ordering, based on the individual orderings.

Now for some conditions on the valuation relation ρ. For every $w \in N$ and every atom p we require that $\rho_w p1$ or $\rho_w p0$ and not both. (This is the Classicality Condition from §6.1. It extends to every formula of $\mathscr{L}$ by induction and guarantees that normal worlds are maximally consistent.) The semantic clauses for the extensional connectives and the modal operators are as in the semantics of §7.2. (In particular, at non-normal worlds ρ relates formulas to truth values directly and is not subject to the Classicality Condition.)

A is commonly believed at w by a group of agents when everyone in the group believes A, believes of everyone in the group that she believes A, and so on. World w is a common belief world of a group

G at world w_1 if everything that's commonly believed at w_1 is true at w. Let's denote w_1's set of common belief worlds by $\mathrm{CB}_G^{w_1}$.

Recall that a soft upgrade with A reorders the worlds so that all the A-worlds are then considered more plausible than all the worlds where A is not true. Formally, following Van Benthem and Liu (2007), the new ordering is defined by the following conditions:

(C1) For all $u, v \in W$, if $\rho_u A1$ and it is not the case that $\rho_v A1$, then $u \,{}_i{\leq}_w^A v$.

(C2) Otherwise, the old ordering remains.

Given we assume a version of (NP), we have some world in the model in which A is true and some world where it's not true. So the upgraded relation will be non-empty. The upgrade also preserves transitivity and well-foundedness. In our modelling, A is going to be part of the explicit content of the fiction.

We can add additional constraints on the upgraded relation: for instance, that (non-normal) worlds which make A true *and* false are less plausible than those making A only true. Or, if the agent follows some pragmatic rules of interpretation, certain worlds obeying those rules might be considered more plausible (for such a pragmatic-based approach, see Bonomi and Zucchi 2003). But for simplicity, let's just stick with the above conditions.

Next we define, for every $S \subseteq W$, the set of most plausible worlds for agent i after the upgrade:

$$\min({}_i{\leq}_w^A, S) = \{v \in S \mid \forall u \in S : v \,{}_i{\leq}_w^A u\}$$

For the corresponding definition for a group of agents G, we simply substitute 'G' for 'i'.

Finally, here's our definition of truth in fiction. Let F be the explicit content of fiction f. For $w \in N$:

(S_1 In f) $\rho_w(\mathrm{In}\, f\, A)1$ iff, for every world $w_1 \in \min({}_G{\leq}_w^F, \mathrm{CB}_G^w)$ and every $B \in F$: $\rho_{w_1} B1$ only if $\rho_{w_1} A1$; and

(S_2 In f) $\rho_w(\mathrm{In}\, f\, A)0$ otherwise.

Thus, A is true in fiction f iff every world considered by group G most plausible with respect to their common beliefs, revised in line with the explicit content F of the fiction, makes A true.

This view has some advantages with respect to Lewis's possible-worlds-only approach, for the most plausible worlds can be impossible. Engaging with *Sylvan's Box* might not require us to revise our ordering much, up to the point where the empty-and-not-empty box appears. Nolan and Hanley say that the most plausible worlds are those where Priest, as the narrator, has false beliefs (relative to the fiction). We take the most plausible worlds as those that comply with the obvious intended reading of *Sylvan's Box*. Priest has true beliefs, relative to the fiction, in which there genuinely is an empty-and-not-empty box. This does not trivialize the theory: the inference from $\mathrm{In}\, f\,(A \wedge \neg A)$ to $\mathrm{In}\, f\, B$ does not go through.

So we can deal with blatant contradictions that are essential to the plot. What about the case of accidental contradictions? Our approach ensures the truth-in-the-fiction of each sentence of the explicit content, even if one contradicts another. But their conjunction need not be true-in-the-fiction. Since we have impossible worlds around, Conjunction Introduction might fail for this particular case. But it need not. It might be that the initial order prefers paraconsistent worlds over all other inconsistent words, in such a way that contradictory sentences generate their inconsistent conjunction, but without explosion. So what a model says about accidental contradictions depends very much on its initial ordering.

This approach faces a couple of worries. The notion of a plausibility ordering is rather vague (just as similarity between worlds is). Our assumption that agents each come equipped with a plausibility ordering has to be viewed as something of an abstraction. Badura (2016) and Badura and Berto (2018) provide responses to this worry. Here, we merely note that the assumption is standard in contemporary epistemic-doxastic logics.

Another objection is that, on our analysis, truth-in-fiction depends too heavily on the community of interpreters. We allow what's true in a given fiction to change over time, as the communal beliefs of

its readers change. Lewis, by contrast, insists that 'what was true in a fiction when it was first told is true in it forevermore' (1978, 44). One might even object that our notion isn't a notion of *truth*, since it fails the governing norm that truth is stable over time (Wright 1992).

On our story, the explicit content of a fiction is treated in an objective, fact-of-the-matter way. One can always check what the author wrote. The worry concerns a fiction's implicit content. Do our interpretations of the fiction aim to discover objective facts about its implicit content, or do they somehow influence that content, as on the model we've presented?

If we go to the literary studies department, we'll always find differing interpretations, with differing claims about what's true in a given fiction. Plausibly, those theorists are each taking something from the fiction's explicit content and evaluating it from the standpoint of their own (shared) knowledge, background, and context. Theirs is a partly creative enterprise, constrained by communal beliefs and norms, as well as by the fiction's explicit content. Differing interpretations of a fiction need not be in strict competition, when given in different historical and social contexts.

Ours is a *contextualist* notion of truth in fiction. An analogy with contextualism about knowledge (DeRose 1992, 2002, Lewis 1996) is helpful. You're in a long queue to post a letter. 'They're open until 6', the person in front of you says, 'and the queue is always shorter then'. What matters to you is whether your informant really knows that the post office is open until 6. For contextualists, that depends on how high the stakes are *for you*, the ascriber of knowledge. If it's vital that you get the letter in the post today then, in your context, there are more possibilities of error floating around and so it's correspondingly harder for you truthfully to ascribe knowledge. It's even harder for the students in the epistemology class to ascribe knowledge to the person in the queue, for they might (for all they know) be brains in vats. In these ways, the truth of 'she knows' may vary with the context of the ascriber.

Our view of 'true in the fiction' is analogous. The truth of '… is true in the fiction' may vary with the context of the interpreter. Lewis

(1978, 44) objects that 'that would mean that what is true in a fiction is constantly changing'. But not so. What changes is the property picked out by 'true in the fiction'. It's not that what the person in the queue knows changes with the importance of your letter (or with how much epistemology you've been reading). What changes is what property 'knows' picks out. Similarly, different communities of interpretation pick out a different property with 'true in the fiction'.

We think the analogy between truth in fiction ascriptions and knowledge ascriptions is more than superficial. 'Knows that *A*' means something like '*A* is true in all contextually uneliminated scenarios' (Lewis 1996, 551). Similarly, '*A* is true in the fiction' roughly means '*A* is true in all the contextually most plausible scenarios'. In both cases, the relevant context is the context of the ascriber (or interpreter). Had the Lewis of 'Elusive Knowledge' (1996) revisited 'Truth In Fiction' (1978), perhaps he would have arrived at a theory similar in spirit to ours?

11.5 Realism and Fictionalism about Fictional Entities

Let's move on to the ontology of fictional objects. (We draw on Berto and Plebani 2015, chapter 13.) Heathcliff is a fictional character created by Emily Brontë. Sherlock Holmes is more famous than any real detective. This is *extra-fictional discourse*: seriously asserted, believed, and seemingly true, in which we seem to refer to fictional characters. We may also quantify over fictional characters, in ways which are difficult to paraphrase away (Van Inwagen 1977):

(11.1) There are characters in some nineteenth-century novels who are presented with a greater wealth of physical detail than is any character in any eighteenth-century novel.

(11.2) Some characters in novels are closely modelled on actual people, while others are wholly products of the literary imagination,

> and it is usually impossible to tell which characters fall into which of these categories by textual analysis alone.

If these are literally true, there must be something in the world that makes them true. *Realist abstractionists* about fictional objects, including Peter Van Inwagen, Saul Kripke, Tatjana von Solodkoff, Amie Thomasson, and Nicholas Wolterstorff, take extra-fictional discourse at face value. According to them, fictional objects are real, existent, abstract entities. That they're abstract explains why we couldn't stumble upon Holmes or Heathcliff.

Some versions of realist abstractionism have it that fictional objects are abstract *productions* (Fontaine and Rahman 2014, Salmon 1998, Thomasson 1999, Voltolini 2006). They are artefacts, brought into reality by the creative activity of authors as they write. Unlike other abstract objects like functions or sets, Heathcliff was created by Brontë and depends on her ontologically: had she not written *Wuthering Heights*, there would be no Heathcliff. Other versions of realist abstractionism differ on this. According to Wolterstorff (1961), Heathcliff exists before Brontë thinks about him: she selects, rather than producing, her characters. Van Inwagen (1977) is non-committal between the creationist and the non-creationist variant.

The creationist version faces issues (Yagisawa 2001). We have some idea of what creation involves for concrete things: what it means for a craftsman to create a chair, or for a mother to create her baby. But it is mysterious what happens when an abstract object is created. Creation seems to involve causation. How can this be, if created abstract objects are devoid of causal features?

One may worry that the creative process looks too fuzzy. Exactly when did Holmes begin to exist? When Conan Doyle set out to write a novel involving a detective? When he wrote down the first sentence of *A Study in Scarlet*? Or perhaps when he wrote the whole first novel, or the first two? Or does Holmes's reality involve collective intentionality? If so, how many readers are needed? Ten or ten thousand? One way to respond to these worries is to point out that creation of physical artefacts is also fuzzy: just when did the table

or the sculpture begin to exist? If there is similar vagueness in both cases, then it can't be a problem specifically for the creationist view of fictional objects.

A more serious worry is that creationists seem unable to get the modal profile of fictional entities right (Jago 2014a, §5.5). If Holmes is a detective, he is inessentially so: he could have been a chemist, or a violinist. He might have done all the things Watson did and vice versa. But if a fictional entity is created by an author's description of them, then it seems that that description will be essential to the character. In writing about a detective living at 221b Baker Street, Conan Doyle creates an entity whose identity is built from those properties. But then, Holmes will be essentially a detective living at 221b Baker Street. That's the wrong result. Of course, Conan Doyle could have written about a detective living elsewhere, or about a violinist. But that character would have been created from a different piece of writing, and so wouldn't have been Holmes.

The non-creationist version of the view faces issues too. Sainsbury (2010) raises the *selection problem*: how could Brontë select one specific abstract entity, rather than another, to be Heathcliffe? We'll discuss the issue in §11.6, where we explore Meinongian theories of fictional characters, which face the exact same issue.

Realist abstractionists of all kinds face a worry with negative existentials: sentences in which we deny the existence of something. Sensible adults deny that Holmes, Heathcliff, or the Big Bad Wolf exist. A father who reassures his son that the Big Bad Wolf doesn't exist aims to speak the strict and literal truth. Negative existentials about fictional objects typically count as extra-fictional discourse, not as intra-fictional ascriptions: in *Wuthering Heights*, Heathcliff very much exists. That's why 'the Big Bad Wolf doesn't exist' needs a treatment along the same lines as 'Heathcliff is one of Brontë's fictional characters'. But realist abstractionists must break this intuitive uniformity: they say, seriously and literally, that Holmes, Heathcliff, and the Big Bad Wolf exist.

Van Inwagen (1977, 308) admits that what to do with 'Mr. Pickwick does not exist' is 'a very complicated question'. Perhaps

these non-existence claims should be read as implicitly restricted quantifications. We are not really claiming that there is no such thing as Heathcliff, on this view, but rather that Heathcliff is not to be found in the realm of *concreta* (Goodman 2004, Thomasson 1999).

We don't think that's plausible. Uses of 'there is' are often restricted, as in, 'there's no wine left'. But restriction is much harder with 'does not exist', which naturally takes the unrestricted reading (Walton 2003). Restricted readings still make sense when made explicit: 'there's no wine left in the house'. But 'Donald Trump does not exist in California' seems nonsense. So we should take 'Heathcliff does not exist' as an unrestricted claim, which realist abstractionists will have to deny.

In the light of all these issues, antirealism about fictional characters becomes attractive. Fictionalism about a discourse (§2.8) takes claims in the area as valuable and useful but not generally true, for the things seemingly talked about are really not there. Talk of fictional objects seems to be a paradigmatic case in point. According to Walton (1990), we play collective games of make-believe when engaging with fiction, involving prescriptions that such-and-such must be pretended.

But as we've already seen, we can't analyse all talk of fictional characters using the fictionalist's 'in the fiction' operator (Van Inwagen 2003). We can't say:

(11.3) In *Wuthering Heights*, Heathcliff is a fictional character due to Brontë.

(11.4) In *Wuthering Heights*, Heathcliff is played by Laurence Olivier in the classic 1939 movie.

The embedded sentences are external to Brontë's fiction, yet still seemingly about Heathcliff. The problem extends to quantification over characters, as in,

(11.5) If no character appears in every novel, then some character is modelled on another character (Van Inwagen 2003, 137).

How can we understand this literal truth, if there literally are no fictional characters?

One option for fictionalists about fiction goes as follows (Jago 2014a, §5.5). Names like 'Heathcliff' surely have a meaning (else we couldn't truthfully say, 'in the fiction, Catherine marries Heathcliff'). So they have a semantic value, to which we can refer and over which we can quantify. But that semantic value isn't a fictional person. 'Heathcliff' doesn't refer to its semantic value. It refers to nothing, since Heathcliff doesn't exist. Brontë's text associates various properties with that semantic value, which can then be accessed by the 'in *Wuthering Heights*' operator.

Extra-fictional discourse is then understood by paraphrase, in terms of the semantic values of fictional names. So, (11.3) says something like: Brontë created the meaning associated with 'Heathcliff'; and (11.4) says something like: Olivier played a role based on the meaning Brontë associated with 'Heathcliff'. Van Inwagen's (11.5) says something like: if no semantic value of a name-in-fiction is shared by all novels, then some semantic value of a name-in-fiction is modelled on some other. Quite a mouthful. No wonder we talk, loosely, in terms of characters! But don't let this loose talk fool you, says the fictionalist. There are no fictional characters.

A third position on the ontology of fictional objects, differing from both realist abstractionism and fictionalism, is the Meinongian option. Let's take a look.

11.6 Non-existent Objects and Impossible Worlds

According to Meinongians (§2.3), fictional characters are non-existent objects. Meinongianism is a form of realism about fictional objects, in that such things are taken as being parts of reality, available for reference and quantification. But, for Meinongians, this is not the same as existing. Meinongians agree with the folk that 'Heathcliff

does not exist' is literally true. No convoluted story about negative existentials is needed.

Non-existent fictional entities straightforwardly accommodate extra-fictional ascriptions, such as 'Heathcliff is a purely fictional character'. But things get more complicated with intra-fictional discourse. To appreciate the issue, we need to say something about the *Comprehension Principle* problem. Meinongians need some 'principle of comprehension' for their objects, telling us which objects are admitted in the theory and what properties they can bear. A naïve version of Meinongianism subscribes to what Parsons (1980) calls the 'Unrestricted Comprehension Principle' for objects:

(UCP) For any condition $A[x]$ with x free, some object satisfies $A[x]$.

(It's unclear whether even Meinong ever endorsed something like this.) Here, '$A[x]$' is a condition such as 'x is a detective, x lives at 221b Baker Street, x is a cocaine addict, x is Moriarty's arch-enemy, …'. (UCP) guarantees that some object is characterized by the condition: call it 'Sherlock Holmes'. Then A[Sherlock Holmes] is literally true. But this can't work, for one can prove anything whatsoever from (UCP) (Priest 2005, xix). Let $A[x]$ be $x = x \wedge B$, with B an arbitrary formula. By the (UCP), something, b, is such that $b = b \wedge B$, from which B follows by Conjunction Elimination.

Nuclear Meinongians like Jacquette (1996), Parsons (1980), and Routley (1980) limit the principle to a restricted vocabulary. They make a distinction between two kinds of predicates, *nuclear* and *extranuclear* (with corresponding properties). Nuclear predicates deal with ordinary features of objects: 'is blue', 'is tall', 'kicked Socrates', 'was kicked by Socrates', 'kicked somebody', 'is golden', and 'is a mountain' (Parsons 1980, 22–3). Extranuclear predicates express logical, ontological, or intentional notions like 'is fictional', 'is possible', 'is thought about by Meinong', 'is consistent' (Parsons 1980, 22–3). Only nuclear predicates are allowed to deliver objects. It is essential that existence be extranuclear, so that one cannot stipulate things into existence by including 'x exists' in a condition.

Both naïve and nuclear Meinongianism entail *literalism* about intra-fictional discourse (Fine 1982). Fictional non-existents like Holmes literally have the properties they are characterized as having. Holmes literally is a detective and literally lives at 221b Baker Street. But this is problematic. Until 2002, 221 Baker Street (there's really no 221b) hosted the Abbey Road Building Society. It has never been the home of any detective. It is literally *false* that that 221 (or 221b) is (or was) Holmes's residence. In one of the stories, Holmes has tea with William Gladstone. But Gladstone never had tea with Holmes, or with any other non-existent entity.

Did Holmes live in a non-existent counterpart of 221(b) Baker Street? Did he have tea with a non-existent counterpart of Gladstone? That seems inconsistent with the data. We want to say that some things in fiction (like Napoleon in *War and Peace*) exist, just as much as we want to say that Holmes doesn't exist. There's no evidence that 'Napoleon' is ambiguous between an existent person and a non-existent fictional character.

Besides, literalism severs intuitive connections between properties. If Holmes really is a detective because he is so characterized (and the feature of being a detective is nuclear), why can't we shake hands with him? We fail to accommodate the insight that if something really is a detective, then it must be a concrete object, with a spatiotemporal address and causal powers. It should, in particular, *exist*. This is, however, denied to Holmes by Meinongians of all kinds.

Dual copula theorists like Zalta (1983, 1988) fare better than nuclear Meinongians. Recall from §2.3 that they propose a distinction between two ways in which things can be ascribed properties. There's ordinary predication, expressing property-instantiation or exemplification, and there's encoding. Encoding a property does not in general entail exemplifying it. Non-existent objects can encode features of any kind (except properties like *encoding F*: see Rapaport 1978). On this view, Holmes encodes, but does not exemplify, *being a detective*. So this view is free from literalism.

Priest (2005) offers a third Meinongian option, with a Qualified Comprehension Principle (see Berto 2012):

(QCP) For any condition $A[x]$, with x free, some object satisfies $A[x]$ *at some world*.

When object o is characterized as $A[x]$, $A[o]$ need not hold at the actual world (though it may). It holds at some world or other. Berto (2008) calls this view *modal Meinongianism*. There is no restriction at all on $A[x]$ in (QCP). And since 'satisfying' expresses ordinary property-instantiation, there is no need for Zalta's dual copula.

The world which realizes the characterized condition may be an impossible world. The modal Meinongian theory needs impossible worlds, just as accounts of imagination (Chapter 7), epistemic and doxastic contents (Chapter 10), and truth in fiction (§11.3) do. With impossible worlds in play, logically, metaphysically, and mathematically impossible contents can all be employed in the characterization of objects, via (QCP).

Modal Meinongianism puts a limit on what Meinong called the Principle of Independence of *Sosein* (the having of properties by objects) from *Sein* (their existential status). For modal Meinongians, some but not all properties are independent from existence. Those which involve the having of causal features, or spatiotemporal location all entail existence (at the actual world). Non-existents like Holmes may bear only those properties that fail to entail existence, such as *being self-identical*, or *being a fictional character*, or *being thought about by Mark on a Monday morning*. But they cannot bear existence-entailing properties, such as *being a detective*, or *being a cocaine addict*.

Crane (2013) proposes a similar framework. He also distinguishes between existence-entailing features, which he connects with Lewisian-Armstrongian natural properties, and non-existence-entailing, 'pleonastic' features. Non-existents like Holmes or Pegasus may instantiate features only of the latter kind and may be represented by intentional agents as instantiating features only of the former kind. Holmes is not a real detective. But he really is a fictional detective and is so insofar as he is represented as a detective by Conan Doyle and his readers.

Modal Meinongianism with (QCP) faces the following issue: what if we characterize something as being *actually* thus-and-so? If we characterize a character as being an *actual detective*, does (QCP) deliver an entity that's a detective at the actual world? Hopefully not, for we would soon fall into absurdity. We could characterize an actual detective just like Holmes, which conflicts with the facts about our world. Or worse still, we could characterize an actual round square, which cannot be part of any possible world.

How can (QCP) avoid this worrying implication? It seems difficult, given that 'actual', 'actually', and 'at the actual world' function by latching on to the actual world, @. Statements embedded within 'actually' always refer back to @ for their evaluation. As a consequence, all truths are necessarily actually true. Obama is necessarily actually the first president of colour (even though he isn't necessarily the first president of colour). 'Actually' is a *rigidifier* for descriptions. So, if 'x is actually F' is to be true at some world, then something must be F in the actual world. Now let our characterizing condition $A[x]$ be of the form: 'actually, $B[x]$'. Then (QCP) entails that something satisfies 'actually, $B[x]$' at some world and hence that something satisfies '$B[x]$' in the actual world. Since this works for any B, we soon run into absurdity.

Impossible worlds to the rescue! The argument above relies on the semantic clause for the 'actually' operator:

(SAct) 'Actually, A' is true at world w iff A is true at @.

Like all semantic clauses we've considered, this should hold of necessity: it applies to all possible worlds w. But it should not apply to impossible worlds. (SAct) tells us that it is impossible for 'actually, A' to be true and for 'A' to fail to be actually true. So there should be an impossible world where 'actually, A' is true, for any A which isn't actually true. So we can infer from 'actually, A' being true at w to A being actually true only when w is a possible world. But the world invoked by (QCP) need not be a possible world. The argument breaks down and the problematic conclusion is blocked (Berto 2012, Priest 2005).

Many realist theories of fictional objects, including non-creationist realist abstractionism and all the Meinongian views we've discussed, face the *selection problem* (Sainsbury 2010), mentioned in §11.5. On all these views, fictional characters exist as abstract or nonexistent entities, independently of the author's intentions and actions. In writing her fiction, she *selects* some of these abstract or nonexistent entities to be her characters. She pins a name on them, but does not endow them with new properties: Holmes was a fictional detective (even though he hadn't appeared in any fiction) before Conan Doyle came along.

How is this achieved? Not through some causal interaction, since we have no causal connection with abstract or non-existent entities. Can an author single out a specific fictional object using a definite description? The problem with this suggestion is that, if a Comprehension Principle delivers Heathcliff, it also delivers other entities extremely similar to Heathcliff. There'll be a whole host of abstract or non-existent entities, differing in (fictional) height by a few millimetres, or in their time of birth by a few seconds, but otherwise alike and as described by Brontë. Which of these is the real Heathcliff? How could Brontë pick just one of them out? Which one is the Heathcliff she refers to?

Perhaps Brontë's Heathcliff is an incomplete object, having no exact height (to the millimetre) and no precise time of birth (to the second)? Perhaps Brontë's Heathcliff has the (fictional) properties she explicitly attributes to him in *Wuthering Heights*, and no more? Heathcliff is (fictionally) 'tall, athletic, well-formed', but is he neither (fictionally) 6ft, nor (fictionally) taller, nor (fictionally) shorter? We do not want to say that it is true in the fiction that Heathcliff is a vague, fuzzy entity with no particular height, weight, time of birth, and so on. In the fiction, he's human, and it's a feature of all humans that they have a precise height at each time.

One bold reply to the selection problem, endorsed by Priest (2016b, chapter 11), has it that we select non-existents via acts of mental pointing. Brontë focuses on Heathcliff through her 'primitive intentionality', singling the character out via pure thought. But it's

hard to see how this can help with the problem of distinguishing between all the candidate Heathcliffs, each 'tall, athletic, well-formed' but differing slightly in their height. Presumably Brontë didn't specify, in her mind, Heathcliff's height to the millimetre. So it isn't clear how she could then intend one rather than another of the candidate Heathcliffs. It could be that 'mental pointing' selects one of the candidates by pointing at random. Then Heathcliff will have (fictional) properties, such as his precise height, which go beyond what we can discover.

An alternative Meinongian response to the selection problem, sketched by Fine (1982) and developed by Berto (2012, chapter 9) and Priest (2016b, chapter 14), drops the realist assumption. On this view, some non-existents are dependent on the mental activities of authors. Heathcliff is available for reference and predication thanks to Brontë's creative skills. Brontë didn't have to select Heathcliff from a pre-established domain. Rather, she somehow produced him and named her product 'Heathcliff'. We then use 'Heathcliff' with the intention of sticking to that original reference.

The view is similar to creationist abstractionism (§11.5). It faces the same problems and seems to add some new ones for good measure. If an author creates something, doesn't it follow that what she created exists? The Oxford English Dictionary has it that 'to create' means 'to bring into existence'. This is fine for realist abstractionists. For Meinongians, though, Heathcliff does not exist. Claiming that one can create non-existents seems to challenge the ordinary meaning of words.

Chapter Summary

We began with the problem of what counts as true in a given fiction, which goes beyond what's explicitly given in that fiction (§11.2). We then considered the problem of inconsistent fictions, which are naturally handled using impossible worlds (§11.3). We presented an account of truth in fiction, which develops one of Lewis's analyses

into an approach which can handle inconsistent fictions with ease (§11.4). We then turned to our second main topic: how we should think about fictional entities. We contrasted realism and fictionalism about fictional characters (§11.5). We found the problems for realism to be serious. We then considered a third option, which takes the Meinongian line that fictional characters are non-existent objects (§11.6). We considered several versions of this idea and their various issues.

12

Counterpossible Conditionals

Co-authored with Rohan French, Graham Priest, and David Ripley

12.1 Why Counterpossibles?

One of the most discussed applications of impossible worlds has to do with the treatment of counterpossible conditionals. These are counterfactuals whose antecedent is true at no possible world. As you may recall from §1.3, the Lewis-Stalnaker semantics has it that, if there are no A-worlds, $A \boxrightarrow B$ comes out automatically true. The conditional with the same antecedent and opposite consequent, $A \boxrightarrow \neg B$, comes out true, too, for the same reason. In general, all counterpossibles are vacuously true. The standard treatment of counterfactuals implies *vacuism* about counterpossibles.

To many, including Brogaard and Salerno (2013), Bernstein (2016), Bjerring (2014), Krakauer (2012), Nolan (1997), and Priest (2008), vacuism seems wrong (§1.3). These authors have come up with numerous examples of counterfactuals with impossible antecedents, such that the consequent matters for the truth value of the whole. Recall Nolan's (1997) pair of Hobbes-sentences from §1.3:

(1.18) If Hobbes had (secretly) squared the circle, sick children in the mountains of South America at the time would have cared.

(1.19) If Hobbes had (secretly) squared the circle, sick children in the mountains of South America at the time would not have cared.

The intuition is that Hobbes's squaring the circle would have made no difference with respect to the life of those sick children. The second Hobbes-sentence should, then, be true for this reason, and for the same reason, the first Hobbes-sentence should be false. (Bernstein (2016) gives a similar argument.)

Other examples of non-vacuous counterpossibles arise with non-causal notions of 'making a difference'. Anna and her singleton, {Anna}, are modally inseparable: necessarily, one exists just in case the other does. Yet we can make good sense of the idea that a particular set's existence depends on a general framework of sets, in a way that Anna's existence doesn't.

(12.1) If there hadn't been any sets, {Anna} wouldn't have existed.

is true, whereas

(12.2) If there hadn't been any sets, Anna wouldn't have existed.

is false. For whether or not sets exist makes no difference to Anna's existence.

Brogaard and Salerno (2013) propose that counterpossibles such as these can help in the analysis of a thing's *essence*. They agree with Fine's (1994) idea that {Anna} is not involved in Anna's essence, even though the two are modally inseparable. They argue that we can explain this using the difference in truth-value between (12.1) and (12.2). We agree that there's a link between essence and counterpossibles such as these, but we're not so sure about Brogaard and Salerno's direction of explanation. Couldn't it be that (12.2) is false because Anna's essence doesn't involve any sets? If so, it may be that essences play a role in explaining counterfactuals (including counterpossibles), and not vice versa.

The rest of this chapter will be largely structured as a dialogue with Timothy Williamson, who presents a series of powerful objections to non-vacuism (Williamson 2007, 2010, 2017). Discussing these objections will give us the opportunity to delve into the details of a non-vacuist theory of counterfactuals with impossible worlds. In §§12.3–12.5, we'll discuss three arguments against non-vacuist semantics and, in §12.6, we'll discuss Williamson's attempts to undermine the intuitive pull of non-vacuism.

12.2 A Semantics for Counterpossibles

The obvious way to free the Lewis-Stalnaker semantics from vacuism is to expand it by adding impossible worlds. Start with a standard propositional language $\mathscr{L}$ like the one of §4.1 and add our counterfactual conditional $\boxright$, so that if A and B are formulas, then so is $A \boxright B$.

A *frame* $\mathscr{F}$ is now a triple $\langle\!\langle W, N, \{R_A \mid A \in \mathscr{L}\}\rangle\!\rangle$, with W the set of worlds, $N \subseteq W$ the subset of normal (possible) worlds, and each R_A an accessibility relation on W (one for each formula in the language). We read '$R_A w w_1$' as meaning that w_1 is *ceteris paribus* like w, but A is true at w_1. (For this reading to make sense, we'll need an extra constraint on each R_A; see below.)

A frame becomes a model $\mathscr{M} = \langle\!\langle W, N, \{R_A \mid A \in \mathscr{L}\}, v\rangle\!\rangle$, when endowed with a valuation function v assigning truth values (0 or 1) to atoms at worlds in N and to all formulas at worlds in $W - N$. (So as before, impossible worlds are worlds where complex formulas are treated as atomic.) The truth conditions for the operators other than $\boxright$ at $w \in N$ are as in §4.1. For simplicity, we do without the accessibility relation for $\Box$ and $\Diamond$, which we treat as unrestricted universal and existential quantifiers over possible worlds. As for the counterfactual:

(S$\boxright$) $v_w(A \boxright B) = 1$ if for all w_1 such that $R_A w w_1$, $v_{w_1} B = 1$, and 0 otherwise.

Logical truth and validity are, respectively, truth and truth preservation at all normal worlds in all models. This gives us classical **S5** modal logic for the extensional connectives and $\Box$ and $\Diamond$. The only operator that looks at impossible worlds is $\Box\!\!\rightarrow$. With no constraints on the accessibility relations R_A, we have a basic system of conditional logic.

Stronger systems can be obtained, as usual, by adding constraints on the accessibility relations. Their intended understanding clearly motivates the following:

(12.3) If $R_A w w_1$ then $v_{w_1}(A) = 1$

(12.4) If $v_w(A) = 1$ then $R_A w w$

The former says that A is true at all R_A-accessible worlds. The latter says that, if A is true at w, then nothing is closer to w than itself. This corresponds to what Lewis (1973b) called 'Weak Centring'.

These conditions have an effect only when w is a possible world, since the R_As are not involved in determining the truth value of anything at an impossible world. They guarantee, respectively, that '$\Box\!\!\rightarrow$' satisfies counterfactual self-implication and *modus ponens*:

(12.5) $\vDash A \Box\!\!\rightarrow A$

(12.6) $A, A \Box\!\!\rightarrow B \vDash B$

These inferences are clearly desirable for the counterfactual conditional.

The semantics is non-vacuist. To see this, consider this model, with $N = \{w\}$:

$$w \xrightarrow[R_{p\wedge\neg p}]{} w_1 \qquad p \wedge \neg p$$

At w, $p \wedge \neg p \Box\!\!\rightarrow q$ is false, even though the antecedent is contradictory. For w can access the impossible world w_1 (via $R_{p\wedge\neg p}$), where q is not true, even though $p \wedge \neg p$ is.

12.3 The Strangeness of Impossibility Condition

In order for the R_As genuinely to express world similarity, we would need to impose a comparative similarity relation on worlds, or a 'system of spheres', expanding Lewis (1973b)'s approach with the addition of impossible worlds. We will not set things up in this way. The problem of how similarity should work when impossible worlds are around is a tricky one. Some of the issues are orthogonal to the topics we are to discuss in this chapter. However, one further constraint on the R_As will play an important role in our discussion, the *Strangeness of Impossibility Condition*:

(SIC) If $v_w A = 1$ for some $w \in N$ and $R_A w w_1$, then $w_1 \in N$.

If A is true at some possible world w, which looks via R_A at w_1, then w_1 is possible, too. The thought expressed by the constraint is the *prima facie* plausible one that, to evaluate the truth at a possible world of a conditional with a possible antecedent, we never look at impossible worlds. Thinking in terms of closeness between worlds, the condition says that any possible world is closer to a possible world w than any impossible world is. Impossible worlds are kept at a distance for as long as they can be: they're *strange*. (Hence the name, due to Nolan (1997). Jago (2014a) and Mares (1997) also endorse the approach.)

With (SIC) in place, it is easily checked that our semantics validates:

(12.7) $\Diamond A, A \boxrightarrow B \vDash \Diamond B$

It has further important consequences for validity, connected to an objection raised by Williamson (2007) over an impossible worlds logic for counterfactuals:

> We may wonder what logic of counterfactuals [non-vacuists] envisage. If they reject elementary principles of the pure logic of counterfactual conditionals, that is an unattractive feature of their position. (Williamson 2007, 174)

Williamson does not make explicit which logic he has in mind as 'the pure logic of counterfactual conditionals', or which of its principles are 'elementary'. But he makes use of a weak counterfactual logic, presented proof-theoretically (2010, 85). We assume that the distinctively counterfactual axioms and rules of this system give a sense of what Williamson means. We will consider three (using '$\vdash$' for theoremhood and '$\leftrightarrow$' for material equivalence):

(12.8) $\vdash A \mathrel{\Box\!\!\rightarrow} A$

(12.9) If $\vdash A \leftrightarrow B$ then $\vdash (A \mathrel{\Box\!\!\rightarrow} C) \leftrightarrow (B \mathrel{\Box\!\!\rightarrow} C)$

(12.10) If $\vdash B_1 \wedge \cdots \wedge B_n \supset C$ then $\vdash (A \mathrel{\Box\!\!\rightarrow} B_1) \wedge \cdots \wedge (A \mathrel{\Box\!\!\rightarrow} B_n) \supset (A \mathrel{\Box\!\!\rightarrow} C)$

Of these, our semantics verifies only (12.8). So should (12.9) and (12.10) be endorsed?

The former has it that whenever A and B are provably equivalent, then so too are $A \mathrel{\Box\!\!\rightarrow} C$ and $B \mathrel{\Box\!\!\rightarrow} C$. If the extensional fragment of the logic is classical (as in our semantics), then any classical contradiction is provably equivalent to any other. So (12.9) implies that $p \wedge \neg p \mathrel{\Box\!\!\rightarrow} C$ is provably equivalent to $q \wedge \neg q \mathrel{\Box\!\!\rightarrow} C$, for any choice of p, q, and C. In particular, $p \wedge \neg p \mathrel{\Box\!\!\rightarrow} q \wedge \neg q$ is provably equivalent to $q \wedge \neg q \mathrel{\Box\!\!\rightarrow} q \wedge \neg q$. But the latter is provable, given (12.9), and hence so is the former:

$$\vdash p \wedge \neg p \mathrel{\Box\!\!\rightarrow} q \wedge \neg q$$

Quite generally, (12.8) and (12.9) imply that any provably contradiction counterfactually implies any other. But why think this? If Graham Priest really had found a box that's both empty and not empty (as in his story, §11.3), would it really be both raining and not raining in Amsterdam? We don't think that's plausible. Since we reject the conclusion, but find (12.8) hard to deny, we reject (12.9).

Similar problems arise in connection with (12.10). Classically we have $\vdash (p \wedge \neg p) \supset q$. From (12.10), we infer

$$\vdash ((p \wedge \neg p) \mathrel{\Box\!\!\rightarrow} (p \wedge \neg p)) \supset ((p \wedge \neg p) \mathrel{\Box\!\!\rightarrow} q)$$

from which, using (12.8), we obtain:

$$\vdash (p \wedge \neg p) \boxrightarrow (p \wedge \neg p)$$

This gives us $\vdash (p \wedge \neg p) \boxrightarrow q$. If we accept all that, then any contradiction will counterfactually imply anything at all. But why think this? It's wrong to think that, were it raining and not raining, giraffes would stand on their horns. So one of (12.8), (12.10), and classical logic must go. We reject (12.10).

Williamson's principles combine to yield bad predictions about counterfactuals with contradictory antecedents. Counterfactual suppositions can take us beyond logical bounds; they can lead us to entertain situations in which logically equivalent claims come apart, or in which a claim can hold without all its consequences holding. For non-vacuists, these are not 'unattractive features' of their view; rather, they provide one of the main intuitive motivations for it. Of course, such intuitions can be challenged: we will come to this in §12.6. But simply assuming that they are wrong would be dialectically unhappy.

Non-vacuists should reject (12.9) and (12.10). There are closely related principles they may accept, however:

(12.11) If $\vdash A \leftrightarrow B$ then $\Diamond A \vdash (A \boxrightarrow C) \leftrightarrow (B \boxrightarrow C)$

(12.12) If $\vdash B_1 \wedge \cdots \wedge B_n \supset C$ then $\Diamond A \vdash (A \boxrightarrow B_1) \wedge \cdots \wedge (A \boxrightarrow B_n) \supset (A \boxrightarrow C)$

These are just like (12.9) and (12.10), except that the validities they yield have as a premise that a certain claim is possible.

Our semantics validates $\Diamond(A \wedge \neg A) \vdash B$ for any A and B, and so the arguments above against (12.9) and (12.10) do not extend to (12.11) and (12.12). With (SIC) in place, we never have to go outside the domain of possible worlds to evaluate an inference, so long as the antecedents of all the conditionals we are dealing with are possible. As a result, all the valid inferences of merely-possible-world semantics (including (12.9) and (12.10)) are recoverable enthymematically, simply by adding suppressed premises of the form $\Diamond A$ (as in (12.11) and (12.12)). In that sense, adding impossible worlds loses us nothing.

We get a lot from accepting (SIC). But is it acceptable? Some authors (Bernstein 2016, Nolan 1997, Vander Laan 2004) have argued against it. In particular, counterexamples have been proposed to (12.7), which follows from (SIC). Nolan (1997, 2017) offers these:

(12.13) If intuitionistic logic came to be thought of as a much more satisfactory basis for mathematics by experts, and if intuitionistic investigations led to breakthroughs in many areas, … then intuitionistic logic would turn out to be correct after all. (Nolan 1997, 550)

(12.14) If Gödel had believed Fermat's Last Theorem to be false, it would have been. (Nolan 1997, 569)

(12.15) If the bag had 63 balls in it, 63 would have been a square number. (Nolan 2017, 17)

Each of these conditionals has the right form to be a counterexample to (12.7): a possible antecedent and an impossible consequent. But are they true? The context for (12.14) involves a person in awe of Gödel's ability, who thinks that whatever was believed by Gödel in mathematics must be true. It seems to us that intuitionistic logic would not turn out to be correct even if most experts agreed on its value, and that Fermat's Last Theorem would stay true, even if Gödel had believed otherwise.

For (12.15), the context is one in which a person teaches a boy how square numbers work by arranging balls in a square grid, then putting them in a bag and counting the balls that come out. Sometimes the total is 16, sometimes 25, and so on. The conditional is then uttered by the person on an occasion where 63 balls are counted. In this case, we agree with Nolan (2017, 17) that (12.15) would be an appropriate thing to say. But it still does not sound literally true to us. Uttering (12.15) in that context seems to us way to convey the thought that some miscounting must have taken place. In that respect, it's like 'if Trump were smart, I'd be a monkey's uncle', whose antecedent is possibly true while its consequent looks

like a metaphysical impossibility. We don't utter such things as a commitment to their literal truth.

One way to motivate (SIC) is by analogy with what Bennett (2003, 227) calls *counterlegal* conditionals. (We thank Jorge Ferreira for calling our attention to this point.) When we evaluate ordinary counterfactuals, we look at worlds like the world of evaluation, up to or around the time of the antecedent, and which are nomologically possible (Bennett 2003, 198). So we already have a 'Strangeness of Nomological Impossibility Condition' in play in the evaluation of ordinary counterfactuals. Nomologically possible worlds form a sphere, in that they are closer to the base world than nomologically impossible ones. So when we evaluate ordinary counterfactuals whose antecedents comply with our laws, we never look beyond nomologically possible antecedent-worlds.

Things are different when we deal with *counterlegals*, whose antecedents are causally or nomologically impossible: 'if gravity obeyed an inverse cube law, then our months would be shorter'. Then we need to move beyond the nomologically possible, and look at the antecedent-worlds that are nomologically most similar to the base world, despite breaking some of its causal laws. (While Bennett doubts that there is any principled way to do it, we are more optimistic. Thought experiments in the natural sciences often have us suppose situations which violate actual physical laws, often with widespread consensus. This suggests that there is a principled way to evaluate the corresponding counterfactual formulations.)

Analogously, we claim that something like (SIC) is in play with counterfactuals whose antecedents do not violate a law which is absolutely necessary. Possible worlds form a sphere, in that they are closer to the base world than impossible ones. We only look at impossible worlds when the antecedent forces us to move outside the sphere of absolute possibility. We do it when engaging in philosophical or logical thought experiments, as when we counterfactually suppose a logical or mathematical theory we deem (necessarily) wrong, in order to draw unpalatable consequences from it, by way of *reductio*. (We will come back to *reductio* reasoning in §12.5.)

12.4 Substitutivity of Identicals

What does our approach to counterpossibles say about identity and the substitution of rigidly coreferential terms? To investigate the issue, we extend our language with n-ary predicates for each n, the two-place identity predicate, '=', and a set of individual constants, with the usual rules of well-formedness. In particular, if 'a' and 'b' are any constants, then '$a = b$' is an atomic formula. (Extending the semantics to the quantifiers is a non-trivial matter, due to the presence of impossible worlds. We need not go into details here, for they are not germane to what follows; but see Priest's 'matrix semantics' (2008, chapters 18 and 23).)

Models now contain a domain and an interpretation function, assigning an element of the domain to each constant, a subset of the domain to each monadic predicate, and (for $n > 1$) an n-tuple to each n-ary predicate. For atomic sentences other than atomic sentence letters and worlds $w \in N$, v_w is defined in the usual way, in terms of the interpretation. In particular, an atomic identity statement '$a = b$' is true iff 'a' and 'b' denote the same element of the domain. (As before, when $w \in W - N$, v_w treats all sentences as atomic.)

As a consequence, v always satisfies these constraints when $w \in N$:

(12.16) $v_w(a = a) = 1$

(12.17) For atomic A, if $v_w(a = b) = 1$ then $v_w(A) = v_w(A[b/a])$

(12.18) For any $w_1 \in N$, $v_{w_1}(a = b) = v_w(a = b)$

It is then easy to establish that, if A is any sentence in which a does not occur within the scope of a '$\Box\!\!\rightarrow$', $w \in N$, and $v_w(a = b) = 1$, then $v_w(A) = 1$ iff $v_w(A[b/a]) = 1$. So the *Substitutivity of Identicals*, as we will call it, holds in such contexts.

As there are no constraints on v at impossible worlds, *Substitutivity of Identicals* does not hold for impossible worlds, just as we would expect. As a consequence, *Substitutivity of Identicals* is not valid when substitution is within the scope of counterfactuals, for counterfactuals

may look to impossible worlds. Again, that's to be expected when impossible antecedents are around. For example,

(12.19) If Hesperus were not Phosphorus, then modern astronomy in particular would be badly mistaken.

That's true, and Hesperus is Phosphorus; yet it's not the case that

(12.20) If Phosphorus were not Phosphorus, then modern astronomy in particular would be badly mistaken.

Rather, it would be (mainstream) modern logic in particular that is badly mistaken.

Now consider the following pair, from Williamson (2007, 174–6):

(12.21) If Hesperus had not been Phosphorus, Phosphorus would not have been Phosphorus.

(12.22) If Hesperus had not been Phosphorus, Hesperus would not have been Phosphorus.

We take the appropriate evaluation of these to be as follows: (12.21) is false and (12.22) is true. Had Hesperus not been Phosphorus, nothing would have followed about the self-identity of Hesperus or Phosphorus. This seems to suggest that surrendering metaphysical truths in a counterfactual supposition does not force us away from logical truths concerning the same subject matter. Our semantics agrees on this. As an instance of (12.5), (12.22) is valid, whereas the fact that Hesperus is Phosphorus does not imply (12.22). In general, on our semantics, $a = b$ does not entail $a \neq b \boxright a \neq a$.

Substitutivity of Identicals can fail on our semantics only when the substitution in question is within the scope of a counterfactual. Counterfactuals create hyperintensional contexts. Our counterfactuals are sensitive to distinctions between impossibilities, which are invisible in a standard intensional framework using possible worlds.

Yet Williamson (2007, 175) finds this 'highly implausible'. The reason for this has two premises. Hyperintensionality, he claims, occurs

only in constructions that are 'about representational features', such as epistemic and intentional contexts. But, he adds, counterfactuals are not about representational features in this way.

There is reason to doubt each of Williamson's premises. First, one might think, with Lycan (2001), that counterfactuals do involve epistemic features (and for reasons wholly independent of the non-vacuism debate). If so, then Williamson's argument falls apart. The failure of substitution in (12.21) and (12.22) would be of a piece with the failure that occurs when we note that it is a priori that Phosphorus is Phosphorus, but not that Hesperus is Phosphorus. (Brogaard and Salerno (2013, 654) appeal directly to the alleged epistemic aspect of counterfactuals to explain failures of substitutivity like this.) One might even agree with Thomasson (2007) that metaphysical modality itself involves representational elements, even when carefully contrasted with epistemic modalities.

That's all rather controversial, and one may not want to commit to any of those views. But even if we decide against them all, we should question the other Williamsonian premise. An operator's being hyperintensional does not entail its being representational or broadly epistemic. According to Nolan (2014), there are hyperintensional contexts that are not 'about representational features', and counterfactuals may well be among these. *Metaphysical grounding* is often taken to be a wholly worldly, non-representational, but hyperintensional concept (Correia and Schnieder 2012, Fine 2012b). If disjuncts ground disjunctive states of affairs or propositions (as many grounding theorists suppose), then *that A* is a ground for *that* $A \vee (A \wedge B)$, but not *vice versa*. So grounding is hyperintensional, since $A \vee (A \wedge B)$ is logically equivalent to A. Wilson (2018) argues that non-vacuism follows from a counterfactual approach to grounding.

Similarly, *essence* is a hyperintensional metaphysical concept which is frequently taken to be wholly non-representational. We can derive 'Anna exists' from '{Anna} exists' and vice versa, and yet what's essential to Anna'a existence differs from what's essential to {Anna}'s existence. In particular, {Anna}'s essence depends on Anna, whereas Anna's essence doesn't depend on {Anna} (Fine 1994). It seems

essential that the state of affairs *that it's raining or not raining* somehow involve the state of affairs *that it's raining*. But that state of affairs isn't essential to *that Trump will be impeached or he won't*. These logically equivalent complex states of affairs have different essences, and yet there's nothing representational or epistemic to them or their essences.

To see how counterfactuals might be hyperintensional without being about representations, simply return to the semantics we sketched above. Assume, together with Barcan, Kripke (1971), and Williamson (2007, 161), that if $a = b$, then it is necessary for a to be b. Notice that our semantics above conforms to this: the truth values of identity statements '$a = b$' do not change across possible worlds. Then a's not being b is a way things just cannot be. Worlds at which a is not b are impossible worlds. There need be nothing epistemic about this, any more than there is about a world which hosts a physical impossibility, such (supposing Einstein was right) as something travelling faster than the speed of light.

One may grant that Williamson's argument about 'representational features' is problematic, but still think that counterfactuals allow for substitution of identicals. Williamson (2007, 174) bolsters this impression with the following argument:

(12.23) If the rocket had continued on its course, it would have hit Hesperus.

(12.24) Hesperus = Phosphorus

(12.25) Therefore, if the rocket had continued on its course, it would have hit Phosphorus.

This, Williamson claims, is 'unproblematically valid' (2007, 174).

We agree that the argument steps are truth-preserving, but deny that this is so in virtue of their logical form alone. The argument isn't logically valid. The steps are truth-preserving (and necessarily so) because the conditional's antecedent is possible (and necessarily so), and this gives the misleading impression that the argument is

formally valid. But the antecedent's being possible is a metaphysical, not a logical fact. To make the argument formally valid, we need to add, as an additional premise, that the rocket's continuing on its course is possible. Then, given (SIC), the argument is formally valid.

12.5 *Reductio* Arguments

Another Williamsonian objection to non-vacuism about counterpossibles comes from *reductio* arguments (Williamson 2007, 2017). These are crucial to mathematics as it is practised. Williamson attempts to show that non-vacuists must hold current standard mathematical practice to be mistaken.

Although Williamson admits that *reductio* arguments need not be formulated in terms of counterfactuals, he takes it as legitimate to do so. And indeed, it is tempting to assert counterfactuals when reporting a particular line of reasoning by *reductio*: 'it can't be that A, because if it were that A, then it would be that B; but B is wrong, so A too must be'. That's valid on our semantics: $A \boxrightarrow B, \neg B \vDash \neg A$. (Weak Centring guarantees that the base world is an A-world if a B-world. But since we check for validity at a possible world, such a $\neg B$-world can't be an A-world, so must be a $\neg A$-world.)

The trouble stems from certain counterpossibles that can be used in *reductio* reasoning in this way. Since the reasoning is good, the counterpossibles ought to come out true. However, Williamson claims that non-vacuists cannot make good on this prediction. He considers the following examples (Williamson 2017):

(12.26) If there were a largest prime p, $p! + 1$ would be prime.

(12.27) If there were a largest prime p, $p! + 1$ would be composite.

(12.28) If there were a largest prime p, $p! + 1$ would be both prime and composite.

Williamson considers the following proof that there is no largest prime. First, establish (12.26) and (12.27) on their own merits, using

standard reasoning. Next, conclude (12.28) from them. Finally, appeal to our knowledge that no number is both prime and composite to conclude that there is no largest prime. As above, the final step of this reasoning is unproblematic for vacuists and non-vacuists alike. The alleged trouble for the non-vacuist is in getting (12.26)–(12.28) to come out true.

Williamson's worry is that a non-vacuist can't appeal to the usual mathematical reasoning we'd use to justify these. We could, ordinarily, reason that $p! + 1$ is not divisible by any $n \leq p$, so (if p is the largest prime) $p! + 1$ has no prime factors and must therefore be prime. However, we're assuming a number of mathematical results here. How can the non-vacuist be sure that they would hold, were there a largest prime? In general, non-vacuists deny that logical entailments carry across to valid counterfactuals: $A \vDash B$ does not imply $\vDash A \mathrel{\square\!\!\rightarrow} B$. So, if p were the largest prime, might it not be the case that $p! + 1$ *is* divisible by some $n \leq p$? How could we be sure either way? But if we don't have access to that reasoning, on the assumption that p is the largest prime, how can we ensure that (12.26) is true?

Similarly, we could ordinarily reason that, if p were the largest prime, then everything greater, including $p! + 1$, would be composite. We could then reason, on this basis, that $p! + 1$ would be both prime and composite. But what entitles the non-vacuist to this reasoning? Perhaps $p! + 1$ would not be greater than p, or perhaps Conjunction Introduction would not be valid, were there a greatest prime.

We think the answer to the puzzle lies in the context sensitivity of counterfactual utterances. Anna can truthfully say 'if I'd hit you, it would have hurt' (because she's got a mean punch); but she can also truthfully say 'if I'd hit you, it wouldn't have hurt' (because Anna wouldn't hurt anyone, so would have punched softly). Suppose, in a friendly conversation with no threat of violence, you ask Anna how strong she is. 'Let's just say', she replies, 'that if I'd hit you right then, it would have hurt'. That seems true (given how strong she is). But now suppose you and Anna are play-fighting. 'Watch it!', you say, as a mock punch comes a little close. 'Don't worry', she says, 'if I'd hit you then, it wouldn't have hurt'. In this way, 'the truth conditions

for counterfactuals … are a highly volatile matter, varying with every shift of context and interest' (Lewis 1973b, 92).

Any broadly Kratzer, Lewis, or Stalnaker-like approach to counterfactuals (Kratzer 1981, 1986, Stalnaker 1968, 1984) essentially involves two ingredients. It has an underlying space of worlds, plus some apparatus for focusing on the ones relevant to interpreting the counterfactual at hand in any particular case. *All* existing approaches to counterfactuals, vacuist and non-vacuist alike, take the second ingredient to be sensitive to the context in which a counterfactual occurs. There is simply no other way to get sensible results.

In the context of *reductio* reasoning, all the usual rules of reasoning must remain available. Similarly, if a counterfactual is uttered in that context, or in the context of reporting on such a proof, then the usual mathematical principles can be called upon to support the counterfactual. In the case of (12.26)–(12.28), in such contexts, conversational participants hold fixed what they know about the additive and multiplicative structure of the natural numbers. With such facts fixed, (12.26) and (12.27) follow easily, with (12.28) following by Conjunction Introduction.

But those mathematical facts need not be held fixed in every conversational context. We might be discussing mathematical finitism (as in Van Bendegem (1994)), and say, quite correctly, that if there had been a greatest number, there would have been a greatest prime number. In that context, we clearly are not retaining the mathematical fact that every number has a successor. Or we might be discussing what the physical world would be like if there were a largest prime number. Again, we cannot allow all of the facts of standard arithmetic to carry over.

As a consequence, valid logical and mathematical reasoning does not automatically carry over into counterfactual reasoning, as a matter of the logic of counterfactuals. *A*'s entailing *B* does not imply that $A \mathrel{\Box\!\!\rightarrow} B$ is valid. But, as we have seen, there may be contexts in which $A \mathrel{\Box\!\!\rightarrow} B$ is true and justifiable on the basis of *A*'s entailing *B*. They include the context of a *reductio* proof, or of explaining or reporting on such a proof. So the non-vacuist can

justify (12.26)–(12.28) and, more generally, she can justify this area of mathematical practice.

It may even be that vacuism about counterfactuals has the most trouble in capturing mathematical practice. The vacuist easily gets the result that (12.26)–(12.28) are true, since she takes all such counterpossibles to be trivially true. But now consider a mathematician explaining principles of constructive mathematics. In the following Q&A, someone raises an objection, 'but given what you're saying, had it been that $\neg\neg A$, then A, and so …'. 'No!' replies the speaker, pointing out that Double-Negation Elimination isn't constructively valid. She rejects as false the counterfactual, 'had constructive mathematics been correct, Double-Negation Elimination would still have been valid'. Her attitude seems to be part of accepted mathematical practice. Vacuists have trouble in accommodating this. (They may take such counterfactuals to be unassertible, or otherwise out of place conversationally. But they can't capture the mathematician's attitude that those counterfactuals are false.)

12.6 Intuitions for Non-Vacuism

We've been defending non-vacuism about counterfactuals from a range of objections. But what are the positive arguments in its favour? Our support for non-vacuism largely rests on our ordinary-language judgements about the truth of a range of counterpossibles, such as the Hobbes-sentences (1.18) and (1.19), and the Anna-{Anna} sentences (12.1) and (12.2). (Jenny (2018) and Nolan (1997) offer further arguments.)

Williamson (2007, 2017) worries about this kind of motivation. He grants that the intuitions behind those judgementsd are present, but argues that they are not veridical. Here, we consider three Williamsonian arguments in this ballpark.

Thinking it Through

The first concerns the following example, due to Nolan (1997). (See also the discussion in Brogaard and Salerno (2013)). Suppose that

you were asked, 'what is 5 + 7?' and answer '12'. Now consider the following sentences:

(12.29) If 5 + 7 were 13, you would have got that sum right.

(12.30) If 5 + 7 were 13 you would have got that sum wrong.

(12.29) seems false and (12.30) true. But if (12.29) really is false then so is vacuism, since it's necessary that 5 + 7 isn't 13. Here is Williamson's response to this case:

> [Such examples] tend to fall apart when thought through. For example, if 5 + 7 were 13 then 5 + 6 would be 12, and so (by another eleven steps) 0 would be 1, so if the number of right answers I gave were 0, the number of right answers I gave would be 1. We prefer (12.30) to (12.29) because the argument for (12.30) is more obvious, but the argument for (12.29) is equally strong. (Williamson 2007, 172, our renumbering)

It seems to us, though, that the argument for (12.29) is not equally strong, for two reasons. First, having concluded that 0 = 1, it proceeds to substitute '1' for '0' within a counterfactual. But in general, that move is invalid (§12.4). Second, whether a particular chain of reasoning succeeds or fails in supporting the truth of a counterfactual depends on context, and in particular on which truths about the case need to be held fixed to legitimate the reasoning (§12.5).

In this case, all we need to hold fixed for (12.30) to be true is that the questioner asked what 5 + 7 is, that the answer given was 12, and that 12 is not 13. Williamson's argument for (12.29) needs to hold fixed all of that, plus facts about decrementing left and right addends. He must assume that 5 + 7 = 13 ⊢ 5 + 6 = 12 and its subtraction-generated cousins remain true, as well as facts connecting 'number of right answers' given to whether someone gets an answer right.

Some counterfactual contexts may support our retaining all of those facts, but not all will. The contexts in which (12.29) comes out true are thus a proper superset of those in which (12.30) comes out true. For to suppose that 5 + 7 is 13 is to suppose that the

additive structure of the numbers is something other than it actually is. Without some special context (such as during a mathematical proof, or reporting on mathematical *reductio* reasoning, as in §12.5), we have reason to expect that we should not hold fixed facts about incrementing and decrementing under such a supposition. So without some special context, we should expect that (12.30) is true whilst (12.29) is not. And it is no good for Williamson to place his argument within one of those special contexts. For as long as there is some context in which (12.29) is false, vacuism is too.

A Heuristic?

Let's go back to our first Hobbes-sentence:

(1.18) If Hobbes had (secretly) squared the circle, sick children in the mountains of South America at the time would have cared.

No matter how we come at this sentence, we find it stubbornly wearing the appearance of a falsehood. Williamson's (2017) explanation for this appeals to a kind of error theory. We naturally take counterfactuals of the form $A \mathrel{\Box\!\!\rightarrow} B$ and $A \mathrel{\Box\!\!\rightarrow} \neg B$ to be contraries: 'if you were to win the lottery you would be happy' and 'if you were to win the lottery you would not be happy' cannot both be true. (Williamson suggests that we may confuse $A \mathrel{\Box\!\!\rightarrow} \neg B$ with $\neg(A \mathrel{\Box\!\!\rightarrow} B)$, thus taking them to be contradictories. But whether or not this is so, contrariety is all his explanation requires.)

This natural tendency is taken as the result of a fallible heuristic for counterfactual conditionals:

(HCC*) If you accept one of $A \mathrel{\Box\!\!\rightarrow} B$ and $A \mathrel{\Box\!\!\rightarrow} \neg B$, reject the other.

(Williamson discusses two potential heuristics, (HCC) and (HCC*), with (HCC) telling us: If B and C are inconsistent, then treat $A \mathrel{\Box\!\!\rightarrow} B$ and $A \mathrel{\Box\!\!\rightarrow} C$ as inconsistent. Williamson prefers to use (HCC*) as it does not make use of the notion of inconsistency. We stick with Williamson's preference here.)

If we evaluate $A \mathrel{\Box\!\!\rightarrow} \neg B$ to be true, then (HCC*) councils that we take $A \mathrel{\Box\!\!\rightarrow} B$ to be false. This is Williamson's explanation of why we take (1.18) to be false (erroneously, in his opinion).

It's unclear whether our reasoning in the case of (1.18) is guided by any such heuristic; and even if it were, it's doubtful that (HCC*) is the right heuristic. It has little plausibility when A is obviously impossible, as Williamson (2017, §6) acknowledges. It's easy to accept both of

(12.31) If it were raining and not raining, it would be raining.

(12.32) If it were raining and not raining, it would not be raining.

contrary to (HCC*)'s advice, for example.

Nevertheless, Williamson (2017, §6) maintains that (HCC*) plays a role when we evaluate a counterpossible to be false. We disagree. Suppose we're asked,

(12.33) If intuitionist logic were correct, would Excluded Middle be valid?

(Imagine we're feigning ignorance, for the benefit of our logic class.) We evaluate by considering situations in which intuitionist logic is correct. We know what these are like, because we understand the principles of intuitionist logic, BHK interpretations, Kripke semantics, and so on. In every such situation, Excluded Middle is not valid. So we judge that

(12.34) If intuitionist logic were correct, then Excluded Middle would not be valid.

is true. But in exactly the same way, we judge that

(12.35) If intuitionist logic were correct, then Excluded Middle would be valid.

is false. The reasoning, via BHK interpretations or Kripke models or whatever, directly leads to our judgement of (12.35)'s falsity. It's not like we first have to work out what we think about (12.34), and then infer which stance to take on (12.35). Williamson's heuristic has nothing to do with it.

Exactly the same goes for Nolan's first Hobbes-sentence,

(1.18) If Hobbes had (secretly) squared the circle, sick children in the mountains of South America at the time would have cared.

According to Williamson, we assesses the conditional by imagining situations in which Hobbes squared the circle. In that situation, it's false that the sick South American children care about Hobbes's achievement. But from there, we can infer, directly, that (1.18) is false. We needn't go via (1.19)'s truth and (HCC*).

Vacuous Quantification

A third argument against non-vacuist intuitions takes the form of an analogy between counterpossibles and vacuous universal quantification. The 'logically unsophisticated', according to Williamson (2007, 173), find it intuitive that 'every golden mountain is a valley' should be false, given that 'every golden mountain is a mountain' is true, on the grounds that *being a mountain* and *being a valley* are incompatible properties. However, both claims are true, vacuously, if there are no golden mountains. People extrapolate wrongly from familiar (non-vacuous) cases.

Williamson (2017, §6) expands the point. We know that dolphins don't have arms or legs and that unicorns have horns, so it's tempting to judge the following as false:

(12.36) Every dolphin in Oxford has arms and legs.

(12.37) Every unicorn is hornless.

Yet these claims are true on the standard treatment of quantifiers, because there are no unicorns or dolphins in Oxford.

The intended analogy with vacuous quantification is clear. The 'logically unsophisticated', such as Franz and Mark, will intuitively judge counterpossibles like (1.18) and (12.2) to be false. But we make the same mistake as in the case of vacuous quantification. Since there are no situations which verify the antecedent, those counterfactuals are true. And, as in the quantification case, contrary pairs of counterfactuals will both be true when there are no situations to verify their antecedents.

But hang on a minute! We'd better not take the analogy too seriously. For by actualist lights, there are no situations verifying the antecedent of any counter-to-fact conditional. If a situation is contrary to fact, then it doesn't exist. If we understood conterfactuals as quantifiers over existing situations, then we'd end up treating them all as material conditionals, with all contrary-to-fact cases coming out trivially true. (One might insist, with Lewis, that there really do exist merely possible situations. But that extreme metaphysical view can't be required to make sense of counterfactuals.)

To make any sense, the analogy to quantification must be situated within a model (or a pretence, or whatever) in which there exist non-actual situations. Then we can take seriously the point that we make mistakes with vacuous quantification. Sure, if there are no situations to verify the antecedent, then we may have a true counterfactual which we're liable to judge as false. So the question is, in general, are there (or should there be) such situations in our best semantic models? In the case of counterpossibles, the question becomes: should there be impossible situations in our semantic models? But this is a key point at issue in the vacuist-non-vacuist debate. It seems dialectically illegitimate to assume that only possible situations may play a role in our models.

Chapter Summary

There are prima facie reasons to think that *vacuism*, the view that all counterpossibles are trivially true, is incorrect (§12.1). We then

offered an impossible worlds semantics for counterfactuals, which makes room for non-trivial counterpossibles (§12.2). The semantics raises a number of questions. One principle which pins down its application is the *Strangeness of Impossibility* condition, which says that, for any given possible world, any impossible worlds is further away from it than any possible world is (§12.3). We discussed a number of Williamson's objections to the non-vacuist approach in the context of (SIC), and argued that they can be overcome.

We then raised the question of whether counterfactuals in general (including counterpossibles) should permit the substitution of rigidly coreferential terms, again by considering Williamson's arguments against non-vacuism (§12.4). The third case that Williamson makes against non-vacuism is that it does not make good sense of the way *reductio* arguments are used in mathematical practice. We showed how non-vacuists can resist this argument (§12.5). Having defended non-vacusim against Williamson's objections, we then considered a range of arguments in its favour (§12.6).

Bibliography

Adams, R. (1974). Theories of actuality, *Noûs* 8: 211–31.

Alechina, N. and Logan, B. (2002). Ascribing beliefs to resource bounded agents, *Proceedings of the First International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS 2002)*, ACM Press, New York, NY, pp. 881–8.

Amit, E., Algom, D., Trope, Y., and Liberman, N. (2009). 'Thou shalt not make unto thee any graven image': the distance dependence of representation, in D. Markman, W. Klein, and J. Suhr (eds.), *Handbook of Imagination and Mental Simulation*, Taylor and Francis, New York, pp. 53–68.

Anderson, A. R. and Belnap, N. D. (1975). *Entailment—the Logic of Relevance and Necessity*, Princeton University Press, Princeton, NJ.

Angell, R. (1977). Three systems of first degree entailment, *Journal of Symbolic Logic* 47: 147.

Armstrong, D. (1989). *A Combinatorial Theory of Possibility*, Cambridge University Press, Cambridge.

Armstrong, D. (1997). *A World of States of Affairs*, Cambridge University Press, Cambridge.

Armstrong, D. (2004). *Truth and Truthmakers*, Cambridge University Press, Cambridge.

Arrow, K. (1950). A difficulty in the concept of social welfare, *Journal of Political Economy* 58: 328–46.

Badura, C. (2016). Truth in fiction via non-standard belief revision, Master's thesis, University of Amsterdam.

Badura, C. and Berto, F. (2018). Truth in fiction, impossible worlds, and belief revision, *Australasian Journal of Philosophy*, doi: 10.1080/00048402.2018.1435698.

Balbes, R. and Dwinger, P. (1975). *Distributive Lattices*, University of Missouri Press, Columbia, MO, and London.

Balcerak Jackson, M. (2016). On imagining, supposing and conceiving, in A. Kind and P. Kung (eds.), *Knowledge through Imagination*, Oxford University Press, Oxford, pp. 42–60.

Baltag, A., Moss, L., and Solecki, S. (1998). The logic of public announcements, common knowledge, and private suspicions, in I. Gilboa (ed.), *Proceedings of TARK 98*, Morgan and Kaufmann, Evanston, IL, pp. 43–56.

Baltag, A. and Renne, B. (2016). Dynamic epistemic logic, in E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, https://plato.stanford.edu/entries/dynamic-epistemic/.

Baltag, A. and Smets, S. (2006). Conditional doxastic models: a qualitative approach to dynamic belief revision, *Electronic Notes in Theoretical Computer Science* 165: 5–21.

Baltag, A. and Smets, S. (2011). Keep changing your beliefs, aiming for the truth, *Erkenntnis* 75: 255–70.

Bar-Hillel, Y. (1964). *Language and Information: Selected Essays on Their Theory and Application*, Addison-Wesley, London.

Bar-Hillel, Y. and Carnap, R. (1953). An outline of a theory of semantic information, *Language and Information: Selected Essays on Their Theory and Application*, Addison-Wesley, Boston, MA, pp. 221–74.

Barker, S. and Jago, M. (2012). Being positive about negative facts, *Philosophy and Phenomenological Research* 85: 117–38.

Barwise, J. and Perry, J. (1983). *Situations and Attitudes*, Bradford Books, MIT Press, Cambridge, MA.

Barwise, J. and Seligman, J. (1997). *Information Flow: The Logic of Distributed Systems*, Cambridge University Press, Cambridge.

Beall, J., Brady, R., Dunn, M., Hazen, A., Mares, E., Meyer, R., Priest, G., Restall, G., Ripley, D., Slaney, J., and Sylvan, R. (2012). On the ternary relation and conditionality, *Journal of Philosophical Logic* 41: 595–612.

Beall, J. and van Fraassen, B. (2003). *Possibilities and Paradox. An Introduction to Modal and Many-Valued Logic*, Oxford University Press, Oxford.

Belnap, N. (1977). A useful four-valued logic, in J. Dunn and G. Epstein (eds.), *Modern Use of Multiple-valued Logic*, D. Reidel, Dordrecht.

Bennett, J. (2003). *A Philosophical Guide to Conditionals*, Oxford University Press, Oxford.

Bennett, K. (2004). Global supervenience and dependence, *Philosophy and Phenomenological Research* 68: 501–29.

Bernstein, S. (2016). Omission impossible, *Philosophical Studies* 173: 2575–89.

Berto, F. (2008). Modal meinoniganism for fictional objects, *Metaphysica* 9: 205–18.

Berto, F. (2010). Impossible worlds and propositions: Against the parity thesis, *The Philosophical Quarterly* 60: 471–86.

Berto, F. (2012). *Existence as a Real Property*, Synthese Library, Springer, Dordrecht.

Berto, F. (2014). On conceiving the inconsistent, *Proceedings of the Aristotelian Society* 114: 103–21.

Berto, F. (2015). A modality called negation, *Mind* 24: 761–93.

Berto, F. (2017). Impossible worlds and the logic of imagination, *Erkenntnis* 82: 1277–97.

Berto, F. (2018). Aboutness in imagination, *Philosophical Studies* 175: 1871–86.

Berto, F. and Jago, M. (2018). Impossible worlds, in E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, https://plato.stanford.edu/entries/impossible-worlds/.

Berto, F. and Plebani, M. (2015). *Ontology and Metaontology: A Contemporary Guide*, Bloomsbury, London.

Berto, F. and Schoonen, T. (2018). Conceivability and possibility: Some dilemmas for humeans, *Synthese* 195: 2697–715.

Bigelow, J. and Pargetter, R. (1990). *Science and Necessity*, Cambridge University Press, Cambridge.

Bjerring, J. (2010). Non-ideal epistemic space, PhD thesis, RSSS, Australian National University.

Bjerring, J. (2012). Impossible worlds and logical omniscience: an impossibility result, *Synthese* 190: 2505–24.

Bjerring, J. (2014). On counterpossibles, *Philosophical Studies* 168: 327–53.

Bjerring, J. and Rasmussen, M. (2017). Two problems of hyperintensionality. Unpublished draft.

Bjerring, J. and Rasmussen, M. (2018). A dynamic solution to the problem of logical omniscience, *Journal of Philosophical Logic*, doi: 10.1007/s10992-018-9473-2.

Blackburn, P., de Rijke, M., and Venema, Y. (2002). *Modal Logic*, Cambridge University Press, Cambridge.

Block, N. (1983). Mental pictures and cognitive science, *The Philosophical Review* 92: 499–541.

Bonomi, A. and Zucchi, S. (2003). A pragmatic framework for truth in fiction, *Dialectica* 57: 103–20.

Brogaard, B. and Salerno, J. (2013). Remarks on counterpossibles, *Synthese* 190: 639–60.

Byrd, M. (1986). Review of Zalta's 'abstract objects', *Journal of Symbolic Logic* 51: 246–8.

Byrne, A. (2007). Possibility and imagination, *Philosophical Perspectives* 21: 125–44.

Byrne, R. (2005). *The Rational Imagination. How People Create Alternatives to Reality*, MIT Press, Cambridge, MA.

Carnap, R. (1947). *Meaning and Necessity*, University of Chicago Press, Chicago, IL.

Carnap, R. (1948). *Introduction to Semantics*, Harvard University Press, Cambridge, MA.

Chalmers, D. (2002a). The components of content, in D. Chalmers (ed.), *Philosophy of Mind: Classical and Contemporary Readings*, Oxford University Press, Oxford, pp. 608–33.

Chalmers, D. (2002b). Does conceivability entail possibility?, in T. Gendler and J. Hawthorne (eds.), *Conceivability and Possibility*, Oxford University Press, Oxford, pp. 145–99.

Chalmers, D. (2002c). On sense and intension, *Noûs* 36: 135–82.

Chalmers, D. (2006). The foundations of two-dimensional semantics, in M. Garcia-Carpintero and J. Macia (eds.), *Two-dimensional Semantics: Foundations and Applications*, Oxford University Press, Oxford.

Chalmers, D. (2010). The nature of epistemic space, in A. Egan and B. Weatherson (eds.), *Epistemic Modality*, Oxford University Press, Oxford, pp. 60–107.

Chellas, B. (1975). Basic conditional logic, *Journal of Philosophical Logic* 4: 133–53.

Church, A. (1936). An unsolvable problem of elementary number theory, *American Journal of Mathematics* 58: 345–63.

Copeland, J. (1979). On when a semantics is not a semantics, *Journal of Philosophical Logic* 8: 399–413.

Correia, F. (2004). Semantics for analytic containment, *Studia Logica* 77: 87–104.

Correia, F. and Schnieder, B. (2012). *Metaphysical Grounding: Understanding the Structure of Reality*, Cambridge University Press, Cambridge.

Costa Leite, A. (2010). Logical properties of imagination, *Abstracta* 6: 103–16.

Crane, T. (2013). *The Objects of Thought*, Oxford University Press, Oxford.

Cresswell, M. (1966). The completeness of S0.5, *Logique et Analyse* 9: 263–6.

Cresswell, M. (1972). The world is everything that is the case, *Australasian Journal of Philosophy* 50: 1–13.

Cresswell, M. (1973). *Logics and Languages*, Methuen, London.

Cresswell, M. (1975). Hyperintensional logic, *Studia Logica* 34: 25–38.

Currie, G. (1990). *The Nature of Fiction*, Cambridge University Press, Cambridge.

Davidson, D. (1965). Theories of meaning in learnable languages, in Y. Bar-Hillel (ed.), *Logic, Methodology and Philosophy of Science*, North-Holland, Amsterdam, pp. 3–17.

Davidson, D. (1970). Mental events, in L. Foster and J. Swanson (eds.), *Experience and Theory*, Duckworth, London, pp. 207–24.

Davidson, D. (1985). *Inquiries into Truth and Interpretation*, Oxford University Press, Oxford.

Dedekind, R. (1901). *Essays on the Theory of Numbers*, Open Court, Chicago, IL.

Dennett, D. (1978). *Brainstorms*, MIT Press, Cambridge, MA.

Dennett, D. C. (1987). *The Intentional Stance*, MIT Press, Cambridge, MA.

DeRose, K. (1992). Contextualism and knowledge attributions, *Philosophy and Phenomenological Research* 52: 913–29.

DeRose, K. (2002). Assertion, knowledge and context, *Philosophical Review* 111: 167–203.

des Rivieres, J. and Levesque, H. J. (1686). The consistency of syntactical treatments of knowledge, *Computational Intelligence* 4: 31–41.

Devlin, K. (1991). *Logic and Information*, Cambridge University Press, New York.

Divers, J. (2002). *Possible Worlds*, Routledge, London.

Divers, J. and Melia, J. (2002). The analytic limit of genuine modal realism, *Mind* 111: 15–36.

Dretske, F. (1970). Epistemic operators, *Journal of Philosophy* 67: 1007–23.

Dretske, F. (1981). The pragmatic dimension of knowledge, *Philosophical Studies* 40: 363–78.

Dretske, F. (2005). The case against closure, in M. Steup and E. Sosa (eds.), *Contemporary Debates in Epistemology*, Blackwell, Oxford, pp. 13–25.

Duc, H. (1997). Reasoning about rational, but not logically omniscient, agents, *Journal of Logic and Computation* 5: 633–48.

Duc, H. N. (1995). Logical omniscience vs. logical ignorance, in C. Pereira and N. Mamede (eds.), *Proceedings of EPIA'95*, LNAI 990, Springer, Dordrecht, pp. 237–48.

Dummett, M. (1973). The justification of deduction, *Proceedings of the British Academy* 59: 3–34.

Dummett, M. (1978a). The justification of deduction, in his *Truth and Other Enigmas*, Harvard University Press, Cambridge, MA, pp. 166–85.

Dummett, M. (1978b). *Truth and Other Enigmas*, Harvard University Press, Cambridge, MA.

Dummett, M. (1993a). *Frege: Philosophy of Language*, Harvard University Press, Cambridge, MA.

Dummett, M. (1993b). *The Seas of Language*, Oxford University Press, Oxford.

Dunn, J. (1976). Intuitive semantics for first-degree entailments and coupled trees, *Philosophical Studies* 29: 149–68.

Dunn, J. (1993). Star and perp: Two treatments of negation, *Philosophical Perspectives* 7: 331–57.

Dunn, J. and Restall, G. (2002). Relevance logic, in D. Gabbay and F. Guenthner (eds.), *Handbook of Philosophical Logic*, 2nd edn, vol. 6, Kluwer Academic, Dordrecht, pp. 1–136.

Duzi, M., Materna, P., and Jespersen, B. (2010). *Procedural Semantics for Hyperintensional Logic*, Springer, Dordrecht.

Eberle, R. (1974). A logic of believing, knowing and inferring, *Synthese* 26: 356–82.

Fagin, R. and Halpern, J. (1988). Belief, awareness and limited reasoning, *Artificial Intelligence* 34: 39–76.

Fagin, R., Halpern, J., Moses, Y., and Vardi, M. (1995). *Reasoning About Knowledge*, MIT Press, Cambridge, MA.

Field, H. (1980). *Science without Numbers*, Princeton University Press, Princeton, NJ.

Field, H. (1989). *Realism, Mathematics and Modality*, Oxford University Press, Oxford.

Fine, K. (1975a). Critical notice: Counterfactuals, *Mind* 84: 451–8.

Fine, K. (1975b). Vagueness, truth and logic, *Synthese* 30: 265–300.

Fine, K. (1982). The problem of non-existents I: internalism, *Topoi* 1: 97–140.

Fine, K. (1994). Essence and modality: The second philosophical perspectives lecture, *Philosophical Perspectives* 8: 1–16.

Fine, K. (2001). The question of realism, *Philosophers' Imprint* 1: 1–30.

Fine, K. (2012a). Counterfactuals without possible worlds, *Journal of Philosophy* 109: 221–46.

Fine, K. (2012b). Guide to ground, in F. Correia and B. Schnieder (eds.), *Metaphysical Grounding: Understanding the Structure of Reality*, Cambridge University Press, Cambridge, pp. 37–80.

Fine, K. (2012c). The pure logic of ground, *Review of Symbolic Logic* 5: 1–25.

Fine, K. (2014). Truth-maker semantics for intuitionistic logic, *Journal of Philosophical Logic* 43: 549–77.

Fine, K. (2016). Angellic content, *Journal of Philosophical Logic* 45: 199–226.

Fine, K. (2017). A theory of truth-conditional content I: Conjunction, disjunction and negation, *Journal of Philosophical Logic* 46: 625–74.

Fine, K. (2019). Constructing the impossible, unpublished article, https://www.academia.edu/11339241/Constructing_the_Impossible.

Fine, K. and Jago, M. (forthcoming). Exact truthmaker logic, forthcoming in *Review of Symbolic Logic*.

Fiocco, M. (2007). Conceivability, imagination and modal knowledge, *Philosophy and Phenomenological Research* 74: 364–80.

Floridi, L. (2015). Semantic conceptions of information, in E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, https://plato.stanford.edu/entries/information-semantic/.

Fodor, J. (1975). *The Language of Thought*, Harvard University Press, Cambridge, MA.

Fontaine, M. and Rahman, S. (2014). Towards a semantics for the artifactual theory of fiction and beyond, *Synthese* 191: 499–516.

Forrest, P. (1986). Ways worlds could be, *Australasian Journal of Philosophy* 64: 15–24.

Frege, G. (1879). *Begriffsschrift, eine der arithmetischen nachgebildete Formelsprache des reinen Denkens*, Louis Nebert, Halle.

Frege, G. (1892). Über sinn und bedeutung, *Zeitschrift für Philosophie und philosophische Kritik* 100: 25–50.

Frege, G. (1956). The thought: A logical inquiry, *Mind* 65: 289–311.

Ganis, G., Thompson, W., and Kosslyn, S. (2004). Brain areas underlying visual mental imagery and visual perception, *Cognitive Brain Research* 20: 226–41.

Gendler, T. (2000). *Thought Experiments: On the Powers and Limits of Imaginary Cases*, Routledge, London.

Gendler, T. and Hawthorne, J. (eds.) (2002). *Conceivability and Possibility*, Oxford University Press, Oxford.

Gioulatou, I. (2016). Hyperintensionality, Master's thesis, Universiteit van Amsterdam.

Goodman, J. (2004). A defense of creationism in fiction, *Grazer Philosophische Studien* 67: 131–55.

Grove, A. (1988). Two modellings for theory change, *Journal of Philosophical Logic* 17: 157–70.

Hanley, R. (2004). As good as it gets: Lewis on truth in fiction, *Australasian Journal of Philosophy* 82: 112–28.

Hawke, P. (2011). Van Inwagen's modal skepticism, *Philosophical Studies* 153: 351–64.

Hawthorne, J. (2004). *Knowledge and Lotteries*, Oxford University Press, New York.

Hawthorne, J. (2005). The case for closure, in M. Steup and E. Sosa (eds.), *Contemporary Debates in Epistemology*, Blackwell, Oxford, pp. 26–43.

Heyd, T. (2006). Understanding and handling unreliable narratives: A pragmatic model and method, *Semiotica* 162: 217–43.

Hill, C. (1997). Imagininability, conceivability, possibility and the mind-body problem, *Philosophical Studies* 87: 61–85.

Hintikka, J. (1962). *Knowledge and Belief: An Introduction to the Logic of the Two Notions*, Cornell University Press, Ithaca, NY.

Hintikka, J. (1975). Impossible possible worlds vindicated, *Journal of Philisophical Logic* 4: 475–84.

Holliday, W. H. (2015). Epistemic closure and epistemic logic I: Relevant alternatives and subjunctivism, *Journal of Philosophical Logic* 44: 1–62.

Horgan, T. (1993). From supervenience to superdupervenience: Meeting the demands of a material world, *Mind* 102: 555–86.

Horgan, T. E. (1982). Supervenience and microphysics, *Pacific Philosophical Quarterly* 63: 29–43.

Hume, D. (1739/1978). *A Treatise of Human Nature*, Oxford University Press, Oxford.

Hyde, D. (1997). From heaps and gaps to heaps of gluts, *Mind* 106: 641–60.

Jacquette, D. (1996). *Meinongian Logic. The Semantics of Existence and Nonexistence*, DeGruyter, Berlin and New York.

Jago, M. (2006). *Logics for resource-bounded agents*, PhD thesis, University of Nottingham.

Jago, M. (2007). Hintikka and Cresswell on logical omniscience, *Logic and Logical Philosophy* 15: 325–54.

Jago, M. (2009a). The conjunction and disjunction theses, *Mind* 118: 411–15.

Jago, M. (2009b). Logical information and epistemic space, *Synthese* 167: 327–41.

Jago, M. (2011). Setting the facts straight, *Journal of Philosophical Logic* 40: 33–54.

Jago, M. (2013a). Against Yagisawa's modal realism, *Analysis* 73: 10–17.

Jago, M. (2013b). The content of deduction, *Journal of Philosophical Logic* 42: 317–34.

Jago, M. (2013c). Impossible worlds, *Noûs* 47: 713–28.

Jago, M. (2013d). Recent work in relevant logic, *Analysis* 73: 526–41.

Jago, M. (2014a). *The Impossible*, Oxford University Press, Oxford.

Jago, M. (2014b). The problem of rational knowledge, *Erkenntnis* 79(6): 1151–68.

Jago, M. (2015). Hyperintensional propositions, *Synthese* 192: 585–601.

Jago, M. (2016). Advanced modalizing problems, *Mind* 125: 627–42.

Jago, M. (2017). Propositions as truthmaker conditions, *Argumenta* 2: 293–308.

Jago, M. (2018a). From nature to grounding, in R. Bliss and G. Priest (eds.), *Reality and its Structure*, Oxford University Press, Oxford, pp. 199–216.

Jago, M. (2018b). *What Truth Is*, Oxford University Press, Oxford.

Jeffrey, R. (1965). *The Logic of Decision*, McGraw-Hill, Chicago, IL.

Jenny, M. (2018). Counterpossibles in science: the case of relative computability, *Noûs* 52(3): 530–60.

Jespersen, B. and Duzi, M. (2015). Introduction to the *Synthese* special issue on hyperintensionality, *Synthese* 192: 525–34.

Kim, J. (1982). Psychophysical supervenience, *Philosophical Studies* 41: 51–70.

Kim, J. (1993). *Supervenience and Mind: Selected Essays*, Cambridge University Press, Cambridge.

Kind, A. (2001). Putting the image back in imagination, *Philosophy and Phenomenological Research* 62: 85–109.

King, J. (1995). Structured propositions and complex predicates, *Noûs* 29: 516–35.

King, J. (1996). Structured propostions and sentence structure, *Journal of Philosophical Logic* 25: 495–521.

King, J. (2007). *The Nature and Structure of Content*, Oxford University Press, Oxford.

King, J. C. (2011). Structured propositions, in E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, http://plato.stanford.edu/entries/propositions-structured/.

Kiourti, I. (2010). *Real impossible worlds: The bounds of possibility*, PhD thesis, University of St Andrews.

Konolige, K. (1986). *A Deduction Model of Belief*, Morgan Kaufman, San Francisco.

Kosslyn, S. and Pomerantz, J. (1977). Imagery, propositions, and the form of internal representations, *Cognitive Psychology* 9: 52–76.

Krakauer, B. (2012). Counterpossibles, PhD thesis, University of Massachusetts.

Kratzer, A. (1981). Partition and revision: the semantics of counterfactuals, *Journal of Philosophical Logic* 10: 201–16.

Kratzer, A. (1986). Conditionals, in A. Farley, P. Farley and K. McCollough (eds.), *Papers from the Parasession on Pragmatics and Grammatical Theory*, Chicago Linguistics Society, Chicago, IL, pp. 115–35.

Kripke, S. (1965). Semantical analysis of modal logic II: Non-normal modal propositional calculi, in J. Addison, L. Henkin and A. Tarski (eds.), *The Theory of Models*, North-Holland, Amsterdam, pp. 206–20.

Kripke, S. (1971). Identity and Necessity, in M. Munitz (ed.), *Identity and Individuation*, New York: New York University Press.

Kripke, S. (1980). *Naming and Necessity*, Blackwell, Oxford.

Kung, P. (2010). Imagining as a guide to possibility, *Philosophy and Phenomenological Research* 81: 620–63.

Ladyman, J. and Ross, D. (2007). *Every Thing Must Go*, Oxford University Press, Oxford.

Langland-Hassan, P. (2016). On choosing what to imagine, in A. Kind and P. Kung (eds.), *Knowledge through Imagination*, Oxford University Press, Oxford, pp. 61–84.

Leitgeb, H. and Segerberg, K. (2007). Dynamic doxastic logic: Why, how, and where to?, *Synthese* 155: 167–90.

Lemmon, E. (1957). New foundations for the Lewis modal systems, *Journal of Symbolic Logic* 22: 176–86.

Levesque, H. J. (1984). A logic of implicit and explicit belief, *Proceedings of the Fourth National Conference on Artificial Intelligence*, Austin, TX, pp. 198–202.

Lewis, C. I. and Langford, C. (1932). *Symbolic Logic*, The Appleton-Century Company, New York.

Lewis, D. (1968). Counterpart theory and quantified modal logic, *Journal of Philosophy* 65: 113–26.

Lewis, D. (1971). Counterparts of persons and their bodies, *Journal of Philosophy* 68: 203–11.

Lewis, D. (1973a). Causation, *Journal of Philosophy* 70: 556–67.

Lewis, D. (1973b). *Counterfactuals*, Blackwell, Oxford.

Lewis, D. (1975). Language and languages, in K. Gunderson (ed.), *Language, Mind and Knowledge*, University of Minnesota Press, Minneapolis, MN, pp. 3–35.

Lewis, D. (1978). Truth in fiction, *American Philosophical Quarterly* 15: 37–46.

Lewis, D. (1981). Ordering semantics and premise semantics for counterfactuals, *Journal of Philosophical Logic* 10: 217–34.

Lewis, D. (1982). Logic for equivocators, *Noûs* 16: 431–41.

Lewis, D. (1983). New work for a theory of universals, *Australasian Journal of Philosophy* 61: 347–77.

Lewis, D. (1986a). Against structural universals, *Australasian Journal of Philosophy* 64: 25–46.

Lewis, D. (1986b). *On the Plurality of Worlds*, Blackwell, Oxford.

Lewis, D. (1994). Humean supervenience debugged, *Mind* 103: 473–90.

Lewis, D. (1996). Elusive knowledge, *Australasian Journal of Philosophy* 74: 549–67.

Lewis, D. (2004). Letters to Priest and Beall, in B. Armour-Garb, J. Beall, and G. Priest (eds.), *The Law of Non-Contradiction—New Philosophical Essays*, Oxford University Press, Oxford, pp. 176–7.

Lycan, W. (1994). *Modality and Meaning*, Kluwer, Dordrecht.

Lycan, W. G. (2001). *Real Conditionals*, Oxford University Press, Oxford.

Macarthur, D. (2008). Pragmatism, metaphysical quietism, and the problem of normativity, *Philosophical Topics* 36: 193–209.

MacColl, H. (1908). 'If' and 'imply', *Mind* 17: 151–2.

McDaniel, K. (2004). Modal realism with overlap, *Australasian Journal of Philosophy* 82: 137–52.

Mackie, J. L. (1977). *Ethics: Inventing Right and Wrong*, Penguin Books, London.

Mares, E. (2004). *Relevant Logic: A Philosophical Interpretation*, Cambridge University Press, Cambridge.

Mares, E. (2010). The nature of information: a relevant approach, *Synthese* 175: 111–32.

Mares, E. D. (1997). Who's afraid of impossible worlds?, *Notre Dame Journal of Formal Logic* 38: 516–26.

Mares, E. D. (2009). General information in relevant logic, *Synthese* 167: 343–62.

Markman, K., Klein, W., and Surh, J. (eds.) (2009). *Handbook of Imagination and Mental Simulation*, Taylor and Francis, New York.

Matravers, D. (2014). *Fiction and Narrative*, Oxford University Press, Oxford.

Meinong, A. (1904). Über gegenstandstheorie, in A. Meinong (ed.), *Untersuchungen zur Gegenstadstheorie und Psychologie*, Barth, Leipzig.

Merricks, T. (2015). *Propositions*, Oxford University Press, Oxford.

Meyer, R. and Van der Hoek, W. (1995). *Epistemic Logic for AI and Computer Science*, Cambridge University Press, Cambridge.

Montague, R. (1970). Universal grammar, *Theoria* 36: 373–98.

Moore, R. and Hendrix, G. (1979). Computational models of belief and the semantics of belief sentences, *Technical Note 187*, SRI International, Menlo Park, CA.

Morreau, M. and Kraus, S. (1998). Syntactical treatments of propositional attitudes, *Artificial Intelligence* 106: 161–77.

Nichols, S. and Stich, S. (2003). *Mindreading. An Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds*, Oxford University Press, Oxford.

Niiniluoto, I. (1985). Imagination and fiction, *Journal of Semantics* 4: 209–22.

Nolan, D. (1997). Impossible worlds: A modest approach, *Notre Dame Journal of Formal Logic* 38: 535–72.

Nolan, D. (2005). *David Lewis*, Acumen, Chesham.

Nolan, D. (2007). A consistent reading of 'Sylvan's Box', *Philosophical Quarterly* 57: 667–73.

Nolan, D. (2013). Impossible worlds, *Philosophy Compass* 8: 360–72.

Nolan, D. (2014). Hyperintensional metaphysics, *Philosophical Studies* 171: 149–60.

Nolan, D. (2016a). Modal fictionalism, in E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, https://plato.stanford.edu/entries/fictionalism-modal/.

Nolan, D. (2016b). The possibilities of history, *Journal of the Philosophy of History* 10: 441–56.

Nolan, D. (2017). Causal counterfactuals and impossible worlds, in H. Beebee, C. Hitchcock, and H. Price (eds.), *Making a Difference*, Oxford University Press, Oxford, pp. 14–32.

Nozick, R. (1981). *Philosophical Explanations*, Clarendon Press, Oxford.

Nute, D. (1975). Counterfactuals and the similarity of words, *Journal of Philosophy* 72: 773–8.

Paivio, A. (1986). *Mental Representation*, Oxford University Press, Oxford.

Parry, W. (1933). Ein axiomensystem für eine neue art von implikation (analitische implikation), *Ergebnisse eines Mathematischen Kolloquiums* 4: 5–6.

Parsons, T. (1980). *Nonexistent Objects*, Yale University Press, New Haven, CT.

Paul, L. A. (2009). Counterfactual theories, in H. Beebee, C. Hitchcock and P. Menzies (eds.), *The Oxford Handbook of Causation*, Oxford University Press, Oxford, pp. 158–84.

Paul, L. A. (2004). Aspect causation, *Journal of Philosophy* 97: 235–56.

Paul, L. A. and Hall, E. J. (2013). *Causation: A User's Guide*, Oxford University Press, Oxford.

Peano, G. (1889). *Arithmetices principia: nova methodo*, Fratres Bocca, Turin.

Pelletier, J. (2003). Vergil and dido, *Dialectica* 57: 191–203.

Pinker, S. (1980). Mental imagery and the third dimension, *Journal of Experimental Psychology* 109: 354–71.

Plantinga, A. (1970). World and essence, *The Philosophical Review* 79: 461–92.

Plantinga, A. (1974). *The Nature of Necessity*, Clarendon Press, Oxford.

Plantinga, A. (1976). Actualism and possible worlds, *Theoria* 42: 139–60.

Priest, G. (1979). Logic of paradox, *Journal of Philosophical Logic* 8: 219–41.

Priest, G. (1987). *In Contradiction: A Study of the Transconsistent*, Martinus Nijhoff, Dordrecht.

Priest, G. (1997a). Editor's introduction, *Notre Dame Journal of Formal Logic* 38: 481–7.

Priest, G. (1997b). Sylvan's Box, *Notre Dame Journal of Formal Logic* 38: 573–82.

Priest, G. (1998). What is so bad about contradictions?, *Journal of Philosophy* 95: 410–26.

Priest, G. (2005). *Towards Non-Being*, Oxford University Press, Oxford.

Priest, G. (2008). *An Introduction to Non-Classical Logic*, Cambridge University Press, Cambridge.

Priest, G. (2014). *One: Being an Investigation Into the Unity of Reality and of its Parts, Including the Singular Object Which is Nothingness*, Oxford University Press, Oxford.

Priest, G. (2016a). Thinking the impossible, *Philosophical Studies* 173: 2649–62.

Priest, G. (2016b). *Towards Non-Being*, 2nd edn, Oxford University Press, Oxford.

Priest, G., Routley, R., and Norman, J. (eds.) (1989). *Paraconsistent Logic: Essays on the Inconsistent*, Philosophia Verlag, Munich.

Proudfoot, D. (2006). Possible worlds semantics and fiction, *Journal of Philosophical Logic* 35: 9–40.

Proudfoot, D. (2018). Sylvan's bottle and other problems, *Australasian Journal of Logic* 15: 95–123.

Pylyshyn, Z. (1973). What the mind's eye tells the mind's brain: A critique of mental imagery, *Psychological Bulletin* 80: 1–25.

Pylyshyn, Z. (1981). The imagery debate: Analogue media versus tacit knowledge, *Psychological Review* 88: 16–45.

Pylyshyn, Z. (2002). Mental imagery: In search of a theory, *Behavioral and Brain Sciences* 25: 157–82.

Quine, W. (1960). *Word and Object*, MIT Press, Cambridge, MA.

Quine, W. (1969). *Ontological Relativity and Other Essays*, Columbia University Press, New York.

Quine, W. V. (1948). On what there is, *Review of Metaphysics* 2: 21–38.

Rantala, V. (1982a). Impossible worlds semantics and logical omniscience, *Acta Philosophica Fennica* 35: 18–24.

Rantala, V. (1982b). Quantified modal logic: Non-normal worlds and propositional attitudes, *Studia Logica* 41: 41–65.

Rapaport, W. (1978). Meinongian theories and a russellian paradox, *Noûs* 12: 153–80.

Rasmussen, M. (2015). Dynamic epistemic logic and logical omniscience, *Logic and Logical Philosophy* 24: 377–99.

Reinert, J. (2018). The truth about impossibility, *Philosophical Quarterly* 68: 307–27.

Rescher, N. and Brandom, R. (1980). *The Logic of Inconsistency. A Study in Non-Standard Possible Worlds Semantics and Ontology*, Blackwell, Oxford.

Restall, G. (1993). Simplified semantics for relevant logics (and some of their rivals), *Journal of Philosophical Logic* 22: 481–511.

Restall, G. (1995). Information flow and relevant logics, in J. Seligman and D. Westerståhl (eds.), *Logic, Language and Computation: The 1994 Moraga Proceedings*, CSLI Press, Stanford, CA, pp. 463–77.

Restall, G. (1996). Truthmakers, entailment and necessity, *Australasian Journal of Philosophy* 74: 331–40.

Restall, G. (1997). Ways things can't be, *Notre Dame Journal of Formal Logic* 38: 583–96.

Restall, G. (1999). Negation in relevant logics (how I stopped worrying and learned to love the Routley star), in D. Gabbay and H. Wansing (eds.), *What Is Negation?*, Kluwer, Dordrecht, pp. 53–76.

Ripley, D. (2012). Structures and circumstances: Two ways to fine-grain propositions, *Synthese* 189: 97–118.

Roca-Royes, S. (2011). Conceivability and *de re* modal knowledge, *Noûs* 45: 22–49.

Rodriguez-Pereyra, G. (2009). The disjunction and conjunction theses, *Mind* 118: 427–43.

Rosen, G. (1990). Modal fictionalism, *Mind* 99: 327–54.

Roush, S. (2010). Closure on skepticism, *Journal of Philosophy* 107: 243–56.

Routley, R. (1980). *Exploring Meinong's Jungle and Beyond*, RSSS, Australian National University, Canberra.

Routley, R. and Meyer, R. (1973). The semantics of entailment I, in H. Leblanc (ed.), *Truth, Syntax, and Semantics*, North-Holland, Amsterdam, pp. 194–243.

Routley, R., Plumwood, V., and Meyer, R. K. (1982). *Relevant Logics and their Rivals*, Ridgeview Publishing Company, Atascadero, CA.

Russell, B. (1903). *The principles of mathematics*, George Allen and Unwin, London.

Russell, B. (1904/1980). Letter to Frege, 12.12.1904, in G. Gabriel, H. Hermes, F. Kambartel, C. Thiel, and A. Veraart (eds.), *Gottlob Frege: Philosophical and Mathematical Correspondence*, Basil Blackwell, Oxford, pp. 166–70.

Sainsbury, M. (2010). *Fiction and Fictionalism*, Routledge, London.

Salmon, N. (1984). Impossible worlds, *Analysis* 44: 114–17.

Salmon, N. (1986). *Frege's Puzzle*, MIT Press, Cambridge, MA.

Salmon, N. (1998). Nonexistence, *Noûs* 32: 277–319.

Salmon, N. (2005). *Metaphysics, Mathematics, and Meaning: Philosophical Papers I*, Oxford University Press, Oxford.

Schacter, D. (1986). Implicit memory: History and current status, *Journal of Experimental Psychology* 13: 501–18.

Schacter, D. and Tulving, E. (1994). What are the memory systems of 1994?, in D. Schacter and E. Tulving (eds.), *Memory Systems 1994*, MIT Press, Cambridge, MA, pp. 1–38.

Schaffer, J. (2008). Knowledge in the image of assertion, *Philosophical Issues* 18: 1–19.

Schaffer, J. (2009). On what grounds what, in D. Chalmers, D. Manley, and R. Wasserman (eds.), *Metametaphysics*, Oxford University Press, Oxford.

Schipper, B. (2015). Awareness, in H. van Ditmarsch, J. Halpern, W. van der Hoek, and B. Kooi (eds.), *Handbook of Epistemic Logic*, College Publications, London, pp. 79–146.

Schwitzgebel, E. (2015). Belief, in E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, http://plato.stanford.edu/entries/belief/.

Scott, D. (1970). Advice on modal logic, in K. Lambert (ed.), *Philosophical Problems in Logic*, Reidel, Dordrecht, pp. 143–73.

Segerberg, K. (1995). Belief revision from the point of view of doxastic logic, *Logic Journal of IGPL* 3: 535–53.

Segerberg, K. (2001). The basic dynamic doxastic logic of AGM, *Frontiers in Belief Revision* 22: 57–84.

Shephard, R. and Metzler, J. (1971). Mental rotation of three-dimensional objects, *Science* 171: 701–3.

Sider, T. (2002). The ersatz pluriverse, *Journal of Philosophy* 99: 279–315.

Sider, T. (2005). Another look at Armstrong's combinatorialism, *Noûs* 39: 679.

Siegel, S. (2006). Which properties are represented in perception?, in T. Gendler and J. Hawthorne (eds.), *Perceptual Experience*, Oxford University Press, Oxford, pp. 481–503.

Siewert, C. (1998). *The Significance of Consciousness*, Princeton University Press, Princeton, NJ.

Sillari, G. (2008). Quantified logic of awareness and impossible possible worlds, *Review of Symbolic Logic* 4: 514–29.

Smiley, T. (1959). Deducibility and entailment, *Proceedings of the Aristotelian Society* 59: 233–54.

Soames, S. (1985). Lost innocence, *Linguistics And Philosophy* 18: 59–72.

Soames, S. (1987). Direct reference, propositional attitudes and semantic content, *Philosophical Topics* 15: 47–87.

Soames, S. (2008). Why propositions cannot be sets of truth-supporting circumstances, *Journal of Philosophical Logic* 37: 267–76.

Stalnaker, R. (1968). A theory of conditionals, in N. Rescher (ed.), *Studies in Logical Theory*, Blackwell, Oxford, pp. 98–112.

Stalnaker, R. (1976a). Possible worlds, *Noûs* 10: 65–75.

Stalnaker, R. (1976b). Propositions, in A. MacKay and D. Merrill (eds.), *Issues in the Philosophy of Language*, Yale University Press, New Haven, CT, pp. 79–91.

Stalnaker, R. (1984). *Inquiry*, MIT Press, Cambridge, MA.

Stalnaker, R. (1991). Indicative conditionals, in F. Jackson (ed.), *Conditionals*, Oxford University Press, Oxford, pp. 136–54.

Stanley, J. (2005). *Knowledge and Practical Interest*, Oxford University Press, Oxford.

Stoljar, D. (2007). Two conceivability arguments compared, *Proceedings of the Aristotelian Society* 107: 27–44.

Tennant, N. (1984). Perfect validity, entailment and paraconsistency, *Studia Logica* 43: 181–200.

Thomas, N. (2014). Mental imagery, in E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, http://plato.stanford.edu/entries/mental-imagery.

Thomasson, A. (1999). *Fiction and Metaphysics*, Cambridge University Press, Cambridge.

Thomasson, A. (2007). Modal normativism and the methods of metaphysics, *Philosophical Topics* 35: 135–60.

Urquhart, A. (1972). Semantics for relevant logics, *Journal of Symbolic Logic* 37: 159–69.

Van Bendegem, J.-P. (1994). Strict finitism as a viable alternative in the foundations of mathematics, *Logique et Analyse* 37: 23–40.

Van Benthem, J. (1979). What is dialectical logic?, *Erkenntnis* 14: 333–47.

Van Benthem, J. (2003). Logic and the dynamics of information, *Minds and Machines* 13: 503–19.

Van Benthem, J. (2011). *Logical Dynamics of Information and Interaction*, Cambridge University Press, Cambridge.

Van Benthem, J. and Liu, F. (2007). Dynamic logic of preference upgrade, *Journal of Applied Non-Classical Logics* 17: 157–82.

Van Benthem, J. and Martinez, M. (2008). The stories of logic and information, in J. van Benthem and P. Adriaans (eds.), *Handbook of the Philosophy of Information*, Elsevier, Amsterdam, pp. 217–280.

Van Ditmarsch, H. (2005). Prolegomena to dynamic logic for belief revision, *Synthese* 147: 229–75.

Van Ditmarsch, H., van der Hoek, W., and Kooi, B. (2008). *Dynamic Epistemic Logic*, Springer, Dordrecht.

Van Fraassen, B. (1969). Facts and tautological entailments, *Journal of Philosophy* 66: 477–87.

Van Inwagen, P. (1977). Creatures of fiction, *American Philosophical Quarterly* 14: 299–308.

Van Inwagen, P. (1998). Modal epistemology, *Philosophical Studies* 92: 67–84.

Van Inwagen, P. (2003). Existence, ontological commitment, and fictional entities, in M. Loux and D. Zimmerman (eds.), *Metaphysics*, Oxford University Press, Oxford, pp. 131–57.

Van Inwagen, P. (2008). McGinn on existence, *The Philosophical Quarterly* 58: 36–58.

Van Leeuwen, N. (2016). The imaginative agent, in A. Kind and P. Kung (eds.), *Knowledge through Imagination*, Oxford University Press, Oxford, pp. 85–109.

Vander Laan, D. (1997). The ontology of impossible worlds, *Notre Dame Journal of Formal Logic* 38: 597–620.

Vander Laan, D. (2004). Counterpossibles and similarities, in F. Jackson and G. Priest (eds.), *Lewisian Themes: The Philosophy of David K. Lewis*, Clarendon Press, Oxford, pp. 258–75.

Varzi, A. (1997). Inconsistency without contradiction, *Notre Dame Journal of Formal Logic* 38: 621–39.

Voltolini, A. (2006). *How Ficta Follow Fiction: A Syncretistic Account of Fictional Entities*, Springer, Dordrecht.

Walton, K. (1990). *Mimesis as Make-Believe: on the Foundations of Representational Arts*, Harvard University Press, Cambridge, MA.

Walton, K. (2003). Restricted quantification, negative existentials, and fiction, *Dialectica* 57: 239–42.

Wansing, H. (1990). A general possible worlds framework for reasoning about knowledge and belief, *Studia Logica* 49: 523–39.

Wansing, H. (2017). Remarks on the logic of imagination: A step towards understanding doxastic control through imagination, *Synthese* 194: 2843–61.

Williamson, T. (1994). *Vagueness*, Routledge, London.

Williamson, T. (1996). Knowing and asserting, *Philosophical Review* 105: 489–523.

Williamson, T. (2000). *Knowledge and Its Limits*, Oxford University Press, Oxford.

Williamson, T. (2007). *The Philosophy of Philosophy*, Blackwell, Oxford.

Williamson, T. (2010). Modal logic within counterfactual logic, in B. Hale and A. Hoffman (eds.), *Modality: Metaphysics, Logic, and Epistemology*, Oxford University Press, Oxford, pp. 81–96.

Williamson, T. (2017). Counterpossibles in semantics and metaphysics, *Argumenta*, 2(2): 195–226.

Wilson, A. (2018). Grounding entails counterpossible non-triviality, *Philosophy and Phenomenological Research* 96: 716–28.

Wittgenstein, L. (1921/1922). *Tractatus Logico-Philosophicus*, Routledge & Kegan Paul, London.

Wittgenstein, L. (1953). *Philosophical Investigations*, Blackwell, Oxford.

Wolterstorff, N. (1961). Referring and existing, *The Philosophical Quarterly* 11: 335–49.

Woodward, R. (2011). Truth in fiction, *Philosophy Compass* 6: 158–67.

Wright, C. (1992). *Truth and Objectivity*, Harvard University Press, Harvard, MA.

Yablo, S. (1993). Is conceivability a guide to possibility?, *Philosophy and Phenomenological Research* 53: 1–42.

Yablo, S. (2014). *Aboutness*, Princeton University Press, Princeton.

Yagisawa, T. (1988). Beyond possible worlds, *Philosophical Studies* 53: 175–204.

Yagisawa, T. (2001). Against creationism in fiction, *Noûs* 35: 153–72.

Yagisawa, T. (2010). *Worlds and Individuals, Possible and Otherwise*, Oxford University Press, Oxford.

Yagisawa, T. (2015). Impossibilia and modally tensed predication, *Acta Analytica* 30: 317–23.

Zalta, E. N. (1983). *Abstract Objects: An Introduction to Axiomatic Metaphysics*, D. Reidel, Dordrecht.

Zalta, E. N. (1988). *Intensional Logic and the Metaphysics of Intentionality*, MIT Press, Cambridge, MA.

Zalta, E. N. (1997). A classically-based theory of impossible worlds, *Notre Dame Journal of Formal Logic* 38: 640–60.

Index